Task relationships and training induced transfer

Joseph Peter Rennie

This dissertation is submitted for the degree of

Doctor of Philosophy

St John's College

University of Cambridge

June 2021

Preface

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the School of Clinical Medicine Degree Committee.

Two chapters of the thesis have already been adapted for publication in peer reviewed journals. In both cases I am the lead author and was responsible for the design, data collection, analysis, and write-up, in collaboration with my supervisory team.

Chapter 2:

Rennie, J. P., Zhang, M., Hawkins, E., Bathelt, J., & Astle, D. E. (2020). Mapping differential responses to cognitive training using machine learning. *Developmental science*, 23(4), e12868.

Chapter 3:

Rennie, J. P., Jones, J., & Astle, D. E. (2021). Training-dependent transfer within a set of nested tasks. *Quarterly Journal of Experimental Psychology*, 2021 Feb, Online ahead of print: https://pubmed.ncbi.nlm.nih.gov/33535924/

Task relationships and training induced transfer

Joseph Peter Rennie

Abstract

This thesis explores how different conceptualisations of task relationships may inform transfer in the context of cognitive training. There are three large empirical chapters: The first uses a novel analysis pipeline applied to a composite of pre-existing datasets, to explore training outcomes across four popular tasks. Specifically, I used two unsupervised machine learning algorithms to identify multivariate task profiles and sub-groups; I then used these to ask whether, and how, task profiles change following training, both across and within sub-groups. The second empirical chapter presents an online cognitive training study exploring transfer patterns within a set of bespoke tasks that are nested hierarchically, and systematically related, according to simple task features. This approach revealed that training at the top of the hierarchy can yield benefits that cascade to lower-level components, but not the reverse. Furthermore, I quantified the overlap between tasks in different ways and then tested which metric best predicts patterns of transfer. In this case, the presence of one particular shared feature across tasks was the best predictor of transfer. In the third and final empirical chapter another large-scale online training study focussed on a set of change detection tasks, again with a nested set of interrelationships. Again, the results speak to the feature specific nature of transfer patterns, but also show that specificity is context dependent. In this case transfer could be bidirectional within the hierarchy. That is, training lower-level components would yield benefits for those same components and in some cases across components, where they appeared within more complex tasks, and likewise, training the complex task would yield benefits in lower-level constituent components. In the final discussion I integrate across these empirical findings, consider how these results fit within the broader cognitive literature and theories of transfer, along with some recommendations for future research.

Acknowledgements

First and foremost, I would like to thank my supervisor Duncan Astle for his mentorship over the past years. His patience and indispensable support allowed me to explore my ideas and develop my skills freely.

Special thanks go to the Medical Research Council and St John's College for their generous financial support and providing me with the opportunity to pursue my research interests. I would also like to thank the many wonderful people I have been fortunate enough to work beside at the Cognition and Brain Sciences Unit for creating such a welcoming and supportive atmosphere. Especially my fellow PhD student cohort and lab members over the years, for all the stimulating discussions, advice, and humour.

Particular thanks go to the following: Mengya Zhang with whom I worked closely on the first project of this thesis, and who more importantly, has showed me great kindness, compassion, and friendship throughout the years; Jonathan Jones, for his advice and input on much of the work presented here; The CALM team and others who let me use their data; Sue Gathercole and Dennis Norris, whose work on cognitive training provided much inspiration for my own; and Becky Gilbert, whose technical support and help in setting up my online studies, was crucial for their success.

Last but not certainly least, I would like to thank all my close friends and family for being so understanding, loving, and supportive throughout, in ways that go far beyond anything I could express with words. I am eternally grateful.

Table of Contents

List of Figures

List of Tables

Chapter 1: General Introduction

1.1 Background and scope

The ability to receive information from the environment and flexibly adapt to it is a hallmark of all living systems (Hasson et al., 2015; Kandel, 2007). Across the lifespan, humans have tremendous capacity to learn and build systems of knowledge, or novel skills, from experience (Karbach et al., 2017; Kievit, 2020; Lovden et al., 2020; Bialystok, 2006; Salthouse & Davis, 2005). The extent to which something learnt in one context generalises to another is of both theoretical and practical importance: theoretically, it provides insight into the spatial and temporal organisation of physiology and cognition; practically, it informs practices in educational and rehabilitative settings, both of which necessitate a carry-over of learning from the original context (Taatgen., 2013; Green & Bavelier., 2008; Barnett & Ceci, 2002).

The field of cognitive training, which formalises the study of cognitive skill acquisition, has been highly controversial. The potential to boost cognitive performance, with the holy grail of transfer to novel tasks, has garnered much interest from academics, educators, clinicians, commercial enterprises, and the public alike (Green & Bavelier., 2008; Simons et al., 2016; Sala & Gobet, 2019). The basic premise being that extended practice – sometimes called training – on one or more cognitive tasks improves performance on other, unpractised, tasks or activities that rely upon shared processes (Taatgen, 2013; Green & Bavelier., 2008; Simons et al., 2016). Cognitive training researchers typically use a set of 'assessment' tasks believed to tap certain aspects of cognition, deployed before and after practice on a different set of 'training' tasks. When practice improves performance on another unpractised task this is taken as evidence of generalisation, implying that something learnt in one context carries over, or 'transfers', to another. The gold-standard design is to compare these effects against an appropriate control condition, to help ensure that they are specifically due to the training itself and not just practice on the assessments themselves (Simons et al., 2016).

The basic tenets of training induced transfer are to be found as far back as Plato's doctrine of formal discipline, in which he posits that the mind contains broad faculties that can be trained and strengthened through activities that engage them in a generalisable manner. Whilst the sentiment may not be new, the formal scientific enquiry of training induced transfer effects dates back just over 100 years, with the main bulk of research being

carried out in the past few decades (Taatgen, 2013; Simons et al., 2016). Whilst the generic pretest-practice-posttest setup has remained fairly consistent across training studies, there has been great variation in study design. The type and number of tasks recruited, practice regimes, participant demographics, control conditions, and inferential procedures, all vary across studies. Indeed, the number of domains over which cognitive training has been examined is vast, ranging from higher-level cognition such as intelligence and reasoning, to intermediate cognitive skills such as Working Memory (WM) and Short-Term Memory (STM), through to lower level processing skills such as visual perception, and motor skills. Researchers have also studied more 'real world' activities, such as: video games, music, chess, math, athletics, and mindfulness.

There was initial optimism that training on one task might yield benefits general enough to carry over to other tasks within the same cognitive domain - sometimes referred to as 'near transfer', and perhaps even beyond - sometimes referred to as 'far transfer' (Green & Bavelier, 2008; Au et al., 2014; Klingberg, 2010; Jaeggi et al., 2008; Holmes et al., 2009). However, after much debate and controversy, evidence is starting to converge: the scope of transfer engendered by typical training protocols is generally limited to 'highly similar' tasks and manifests in task specific processes (Melby-Lervag et al., 2016; Sala & Gobet, 2019; Simons et al., 2016; Soveri et al., 2017; Gathercole et al., 2019). That is, training improvements rarely (if ever) transfer to other task domains, often fail to transfer even within-domain, and in some cases fail to transfer to tasks that differ by only a single feature but otherwise identical. Many of the early mixed findings and interpretations in the field have since been shown, amongst other things, to be due to methodological shortcomings such as: lacking statistical power, failing to correct for multiple comparisons, or lacking an appropriate control group (Sala & Gobet, 2019; Simons et al., 2016).

To explain the limited scope of transfer and better understand its boundary conditions, researchers have called for a more systematic approach to cognitive training research (Katz et al., 2017; Redick, 2019; Sala & Gobet, 2019; Smid et al., 2020; Von Bastian & Oberauer, 2014; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Taatgen, 2013). Towards this aim, the current thesis explores how different ways of conceptualising task relationships can inform training induced transfer effects and their interpretation. A full coverage of the training literature is beyond the scope of this current thesis, instead I focus on what would classically be described as the sub-fields of working memory (WM), short term memory (STM), visual perception, and attention. However, for the purposes of this general

introduction I refrain from including domain specific reviews and instead provide a more general overview. I provide more specific reviews in each experimental chapter that are relevant to the tasks therein. My purpose in structuring it this way (domain-general overview, followed by a domain-specific review within each chapter) is to allow me to discuss two crucial broader issues here, namely task impurity, the extent to which any given task measures an intended construct, and task similarity, the extent to which two tasks overlap (see the 'Task similarity and transfer' section for more details).

1.2 Theories of between-task transfer

Transfer holds important theoretical implications for models of skill acquisition and performance (Singley & Anderson, 1989; Barnett & Ceci, 2002). Modern accounts of transfer are either rooted in, or closely related to, production system models (Anderson, 1982; Cole et al., 2012; Gathercole et al., 2019; Singley & Anderson, 1985; Newell, 1990; Taatgen, 2013), in which, task performance is achieved by stimulus information being inputted to, and propagated via, a series of processing components (production rules), to produce an output. These processing components are functions that take information from the senses and/or current memory state as input, and pass them to a set of conditional statements, each of which specifies an output that either modifies the memory state or initiates a motor response.

Learning in the context of production system models concerns the acquisition, modification, and composition of these processing components. This is thought to follow a declarative-to-procedural trajectory – the component processes used to perform a task start out very general and inefficient but with experience become increasingly specialised and efficient (Taatgen, 2013). Lower-level processing components are combined sequentially into 'modules' to form sub-routines within a task-routine (sometimes referred to as a task-set, Rogers & Monsell, 1995), and when these sub-routines can be used effectively by other task-routines there is potential for transfer (Taatgen, 2013; Gathercole et al., 2019). From this perspective, transfer varies continuously according to the relative utility and interchangeability of modules at any given point in time. This provides a nuanced and dynamic interpretation of transfer, how and when it appears, and how it might vary across different stages of development and with varying amounts of practice. Moreover, it provides a way of deriving a taxonomy of tasks according to the interchangeability of their sub-routines. In turn this provides a concrete way of defining the overlap between tasks, a prerequisite for making quantitative predictions about transfer (Reder & Klatzky, 1994; Barnett & Ceci, 2002; Taatgen, 2013; Gathercole et al., 2019).

At the heart of this perspective is a division between task-specific and task-general processes, with the latter being essential for transfer (Singley & Anderson 1985, Taatgen, 2013). The initial optimism of cognitive training research reflected the prospect that training might improve task-general processes that are shared across many tasks, within and even between cognitive domains (Klingberg, 2010). However, more recent investigations have demonstrated the feature-specificity of transfer (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Sala & Gobet, 2019; Soveri et al., 2017). For example, while transfer has been observed between n-back variants using different stimulus types (Holmes et al., 2019; Minear et al., 2016; Soveri et al., 2017; Waris et al., 2015), complex span training effects appear tied to the specific stimulus type (Holmes et al., 2019; Minear et al., 2016). Furthermore, digit span training does not readily transfer to other simple span tasks that differ by only a single stimulus feature, such as modality (visual vs auditory presentation of digits) or stimulus type (e.g. digits vs letters; Norris et al., 2019).

Gathercole et al. (2019) propose a framework in which transfer varies not as a function of similarity with respect to all the shared processing components between tasks, but instead primarily as a function of the applicability of *novel cognitive routines*, acquired during training, to an untrained task. A useful cognitive routine acquired during training can be conceptualised as a higher order process that controls the flow of lower order processes in a novel manner, in order to facilitate performance (e.g. mnemonics, chunking, and proactive control strategies; Gathercole et al., 2019; Taatgen, 2013). Acquiring new cognitive routines is resource intensive, so it is likely that we will only develop new ones if they improve performance in a meaningful way. Gathercole et al. (2019) suggest that the cognitive routines recruited to perform some tasks are relatively well established and functional, so the development of new ones is not necessary. Moreover, when people do develop new routines to enhance performance after extensive practice, these tend to be tied to the specifics of the stimuli and/or paradigm and thus do not readily transfer. Accordingly, two tasks may be relatively highly correlated but show no transfer to one another following training. On the other hand, when task demands are relatively novel, they require the acquisition of more rudimentary routines that are less tied to the specifics of the task and thus transfer more readily. This framework imposes important theoretical constraints on transfer by emphasising the role *novel* task demands play in necessitating the development of new higher order routines, and their shared utility across different tasks.

Cognitive routines are closely related to the concept of a ‘task-set’, introduced by Rogers & Monsell (1995). It too describes the set of processes used by an individual to link sensory input to motor output to accomplish a task. According to Rogers & Monsell (1995), task-sets can be adopted in a preparatory manner so as to form an ‘effective intention’ to perform a task, and can be brought about both endogenously (e.g. proactive conscious preparation) and exogenously (e.g. in reaction to an external stimuli). One possibility is that task-sets perform a shielding function helping to prevent irrelevant stimulus features from interfering with response processes (Rogers & Monsell, 1995; Dreisbach & Wenke, 2011). Whilst task-sets may be adaptive in most contexts, they may also engender task specificity and even negative transfer in training contexts, as well as switch costs within task-switching contexts.

1.3 Task similarity and transfer

What exactly does similarity mean in the context of a cognitive task? Despite an increasing convergence of evidence suggesting transfer of improvements is largely constrained to tasks that are ‘highly similar’ to those being trained, there is still no commonly agreed upon method for operationalising task similarity and consequently no taxonomy by which to determine how ‘near’ or ‘far’ two tasks are, making it difficult to establish the precise boundary conditions for transfer (Gathercole et al., 2019; Taatgen, 2013). Researchers rarely formalise task similarities beyond classical interpretations of what the tasks are purported to measure. Operationalising task similarity is a non-trivial and fundamental issue in the field of cognitive training and indeed cognitive science more generally (Barnett & Ceci, 2002; Kievit et al., 2011; Maul et al., 2016; Meyer et al., 2001; Miyake et al., 2000; Taatgen, 2013). Nonetheless, there exists several ways of operationalising task similarity to date, each of which has its associated pros and cons.

A common approach is to define the similarity between tasks according to their correlational properties (behaviourally or physiologically). Correlations provide a convenient quantitative measure of the linear relationships between tasks and can also be used in combination to represent cognitive abilities at the latent level (Gogtay & Thatte, 2017; Miyake et al., 2000; Loehlin, 1987). Multivariate approaches to cognitive training such as Structural Equation Modelling (SEM) provide a potentially useful tool for identifying whether training has impacted cognition at the level of a cognitive domain. They allow researchers to examine changes in constructs representative of latent abilities following training, something that is otherwise difficult to ascertain by comparing changes on

individual tasks alone, as these changes could stem from multiple processes both specific and general (Schmiedek et al., 2010; Karbach et al., 2017; Protzko, 2017; Taatgen, 2013; Smid et al., 2020). However, neither the correlations themselves, nor the latent constructs commonly used by psychologists to group tasks, are predictive of transfer, and both are subject to change as a function of experience developmentally and as a consequence of training (Gathercole et al., 2019; Schmiedek et al., 2010; Smid et al., 2020; Kievit, 2020). Moreover, they alone cannot be used to make causal inferences, or tell us about underlying mechanisms, and must be combined with theory and predictive models in order to do so. Further, there is a danger of circular reasoning in factor analytical approaches – so called 'observed variables' are often labelled according to the latent factors to which they have been previously associated, creating an illusion of theoretical reasoning where there is none (Kievit et al., 2011; Maul et al., 2016).

Another approach is to organise tasks according to the composition of hypothetical cognitive modules, or processing components, such as in the production or connectionist style models described in the above section (Anderson, 1982; Feldman & Ballard, 1982; Singley & Anderson, 1985; Taatgen, 2013; Yang et al,. 2019; Smid et al., 2020). Researches may specify or generate (e.g. using a reinforcement learning algorithm) a series of processing components that take information from either the senses (i.e. stimulus information), a memory state, or both as input, and give an output that either modifies the memory state, or produces a response. As mentioned previously, this allows researchers to specify underlying mechanisms and define task relationships, and thus task similarity, according to the interchangeability of their processing components (Taatgen, 2013). However, generating such models is often labour intensive (especially with production models), and requires many assumptions and abstractions, which in themselves require extensive experience, as well as theoretical and technical knowledge (about which there is rarely a consensus). That is, the ability to specify hypothetical processing components used to perform tasks in a way that has any chance of meaningfully representing, reproducing, or explaining 'real-world' observations, depends upon one's ability to sensibly constrain both the functional aspects of the components and their learning parameters, which in turn depends upon experience with real-world observations, mathematical abstractions of them, and any software/programming protocols used to implement them. Ironically perhaps, from a productionist's perspective, the ability of one information processing system (i.e. a human), determines the ability of another information processing system (i.e. a computational model), to determine the ability of the

former. As such, these types of modelling approaches are not readily available for many researchers, although this may start to change with the increasing availability of modern machine learning software packages.

Ultimately, whichever approach researchers choose, relies upon the identification, specification, and variation of the extrinsic task-features (e.g. stimulus type, spatial properties, timings, and goals) from which they are comprised (i.e. task analysis). This also provides an approach to operationalising task similarity in and of itself (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Soveri et al., 2017). Importantly, these correlational, theoretical, and task analytical approaches to operationalising task similarity are not mutually exclusive, in fact they are most likely mutually reinforcing, and presumably map onto one another in some fashion. However, the field is now calling for more systematic and tightly controlled manipulations of extrinsic task features in high powered studies, in order to better establish the boundary conditions of transfer and inform cognitive theory (Katz et al., 2018; Redick, 2019; Sala & Gobet, 2019; Von Bastian & Oberauer, 2014; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Smid et al., 2020; Taatgen, 2013). As such, much of the approach taken, and language used in this thesis (particularly in Chapters 2 and 3), represents a conceptual shift away from organising tasks by the hypothetical constructs that they purportedly tap, and towards organising them instead by the extrinsic task-features from which they are comprised.

1.4 Individual differences in training and transfer

Understanding the role of individual differences in training outcomes may also help explain some of the inconsistencies in the literature and has thus received increasing attention from researchers (Smid et al., 2020). The longest-standing example of this is the aptitude by treatment interaction (e.g. Cronbach, 1957; Ferguson, 1956; Snow, 1989), or in other words, how an individual's current cognitive ability interacts with their training outcome. Two popular accounts have emerged, namely: the compensation account and the magnification account (Lovden, Brehmer, & Lindenberger, 2012). The compensation account suggests that those with higher baseline scores have less to gain, as they are closer to ceiling prior to training; this assumes that there is plateau in overall performance, with some subjects being closer to this before they start training. Conversely, the magnification account suggests that those with higher baseline scores will show greater improvements, because they have more cognitive resources available in order to maximise on the potential benefit of the training – e.g. to develop strategies (See Karbach, Konen, & Spenglar, 2018; for a recent overview).

One proposal for predicting the conditions under which either compensation or magnification effects arise, is to make a distinction between the flexibility vs the plasticity of cognition (Lovden et al., 2010; Lovden et al., 2012). Here, flexibility denotes the capacity to optimise task performance within a set of currently available cognitive states. That is, to adapt flexibly to the ongoing environmental demands by assuming certain states that are already available. On the other hand, plasticity denotes the capacity to change the range and/or content of the available cognitive states to optimise task performance. That is, to adapt to the ongoing environmental demands by modifying or expanding the repertoire of available states. According to Lovden et al (2012), performance gains acquired primarily via flexibility are likely to show transfer patterns consistent with a compensation account, whereas performance gains acquired primarily via plasticity are likely to show transfer patterns consistent with a magnification account. They argue that if the brain is already optimised to perform a task within its current structural constraints, then it has nothing to gain from altering the way the task is executed, hence why, within the range of performance covered by flexibility, individuals with high performance have less to gain from practice. In contrast, if practice pushes individuals beyond the current range of performance, inducing plastic changes in the brain, then within the range of performance covered by plasticity, individuals with high performance have more to gain because presumably their baseline scores are at least in part a reflection of their previously manifested plasticity. This line of reasoning echoes the sentiments expressed by Gathercole et al (2019) concerning the importance of novelty for acquiring new cognitive routines.

It is well documented that both the flexibility and plasticity of cognition, along with their corresponding physiological substrates, change dynamically across the lifespan (Hertzog et al., 2008; Lovden et al., 2020; Bialystok et al., 2006; Salthouse & Davis, 2006; Kievit, 2020). As such, the aptitude by treatment interaction is closely related to, and may help explain, some of the age-related differences found in training and transfer. Indeed, some have found training and transfer effects in one age bracket that are not present in another (Bürki et al., 2014; Karbach et al., 2017 ; Gathercole et al., 2019; Shmiedek et al., 2010; Borella et al., 2014). For example, Borella et al (2014) found that following WM training, younger participants who performed relatively well on a measure of crystalized intelligence saw greater transfer to other measures of WM, inhibition, and reasoning (magnification). In contrast, they found that older participants who performed relatively poorly on measures of crystalized intelligence and WM saw greater transfer to other WM and STM tasks

(compensation). Similarly, Karbach et al (2017) trained participants across a range of ages (children, younger adults, and older adults) on a task switching paradigm and found that both the younger and older adults who performed worse at baseline showed greater transfer to a similar but untrained task switching paradigm, thereby reducing age related differences, and showing evidence in line with a compensation account. Naturally, these aptitude by treatments effects may also extend to clinical populations and help toward tailoring interventions to meet the needs of the individual (Karbach et al., 2017). Whilst the extreme accounts of magnification and compensation are likely oversimplifications and will undoubtedly need modifying, they nonetheless provide a useful starting point for explaining discrepancies in training and transfer effects due to individual differences (Borella et al., 2014; Karbach et al., 2017; Smolen, Jastrzebski, Estrada, & Chuderski, 2018).

Other individual differences too may prove useful in understanding the inconsistencies in training and transfer effects across studies, and in turn help design more effective training interventions. Intrinsic motivation, progressive difficulty and purposeful practice are all important pre-requisites to effective learning and skill acquisition (Bjork & Bjork, 2014; Campitelli & Gobet, 2011; Cordova and Leper, 1996; Green and Bavelier, 2008; Hertzog et al., 2008; Jaeggi et al., 2014; Smid et al., 2020). Participants' subjective reports of intrinsic motivation during training indicate that those reporting greater intrinsic motivation tend to show enhanced training gains and transfer effects (Green and Bavelier, 2008; Hertzog et al., 2008; Jaeggi et al., 2014; Mohammed et al., 2017). Participants report feeling more motivated and engaged when training tasks are 'gamified', for example by providing level ups, performance feedback, and framing/displaying tasks in a fantasy/game like narrative (Adams, 2014; Green and Bavelier, 2008; Mohammed et al., 2017; Lumsden et al., 2016). On the other hand, providing extrinsic motivation by the way of monetary rewards appears to have relatively little effect upon training outcomes (Jaeggi et al., 2014; Lumsden et al., 2016). Furthermore, building incremental improvement structures into training tasks, allowing for their difficulty to be frequently updated so that it is just beyond a participant's current level of competency (adaptive training), and providing relevant feedback, allows the user plenty of opportunities to employ 'purposeful practice' and adjust their performance accordingly (Bjork & Bjork, 2014; Cordova and Leper, 1996; Green and Bavelier, 2008; Campitelli & Gobet, 2011; Jaeggi et al., 2014; Lumsden et al., 2016). Indeed, studies using adaptive training regimes and feedback have shown to improve training outcomes above and beyond those that do not (Jaeggi et al., 2014; Green and Bavelier, 2008; Lumsden et al., 2016).

Tangentially, cognition-related beliefs such as an individual's belief about the malleability of intelligence also appear to contribute to the effects of training, with those who believe intelligence to be more malleable showing greater training outcomes (Jaeggi et al., 2014).

1.5 Summary

Training induced transfer effects carry important implications for our understanding of cognition, as well as for a range of learning and rehabilitative settings. There is an increasing convergence both theoretically and empirically to suggest that the transfer engendered by typical training studies is tied to specific task features (Melby-Lervag et al., 2016; Sala & Gobet, 2019; Simons et al., 2016; Gathercole et al., 2019; Norris et al., 2019). As such, higher powered studies, that systematically manipulate features in a tightly controlled fashion across both the training and assessment tasks, are required to help further identify and understand the precise boundary conditions of transfer (Holmes et al., 2019; Norris et al., 2019). Aside from more systematic experimental designs, there is also a need for new analytical methodologies and lines of exploration to move beyond univariate comparisons at the group level and toward multivariate comparisons that may also allow for a better understanding of the role individual differences have play in training outcomes (Jaeggi et al., 2014; Karbach et al., 2017; Smid et al., 2020).

1.6 Aims and structure of the current thesis

The overarching aim of the current thesis is to explore how different conceptualisations of task relationships inform training induced transfer effects and their interpretation.

Chapter 1 of the thesis explores the use of a novel methodological alternative for investigating individual differences in responses to training and the effects of training more generally. Specifically, I used two simple unsupervised machine learning algorithms, namely: self-organising-maps (SOMs) and K-means-clustering. The SOM algorithm is essentially a non-linear data reduction technique that allows multivariate data to be organised and represented topologically. I used the SOM algorithm to model multivariate data across four popular WM tasks in a large composite dataset. I then used the K-means clustering algorithm to identify four distinct performance profiles within the SOM model. I then asked questions about the effects of training, using a separate composite dataset comprised of children from various studies who had undergone WM training and for whom there was data on the four WM tasks at pre- and post-training. First, I tested whether the way in which tasks are

represented by the SOM model changes following training. Second, children who had undergone training were allocated to one of the four performance profiles identified by the K-means clustering. I then asked whether children moved to a different group following training and whether changes in group membership predicted scores on a separate measure of fluid intelligence. Importantly, this approach provides an alternative for analysing cognitive training data that goes beyond changes in individual tasks and instead looks at different changes in the relationships across tasks and individuals.

Chapters 2 and 3 of the thesis are experimental chapters with new data collection. Both are large online training studies. Each is the result of a large amount of pilot experimentation, to first check task designs and difficulty. Here, for brevity, I present the final fully powered versions of each study, using adaptive training regimes, powered to detect small-medium effect sizes, randomised group allocation, and active control conditions. Importantly, both utilised sets of tasks for assessment and training that varied systematically with respect to their extrinsic task features. The tasks were hierarchically nested with respect to their combination of task features. That is, the higher-level tasks contain *all the features* of their lower-level counterparts. This approach allowed task overlap to be quantified in an unambiguous manner, which in turn allowed me to establish specific extrinsic task features as potential boundary conditions. Moreover, the hierarchical nature of the task sets allowed me to ask questions about complexity and the direction of transfer cascades.

Specifically, the experiment presented in Chapter 2 explored the potential transfer effects of training on two tasks within a set of six hierarchically nested perceptual discrimination tasks. Whilst the stimulus set (spikey 2D shapes) was identical across tasks, the task features varied systematically with respect to judgement type (number of spikes or ‘spikiness’), presentation type (simultaneous or delayed), and task-switching, allowing me to establish them as potential boundary conditions. All participants completed each of the six assessment tasks both before and after training, one group received training on a relatively low-level task, another group received training on a relatively high-level task, and a third group trained on a control task. This design allowed me to ask whether training on different parts of the hierarchy produced different transfer patterns and whether these transfer patterns were predicted by different metrics of task similarity.

Similar in its conception, the experiment presented in Chapter 3 explored the potential transfer effects of training on a set three hierarchically nested change-detection-tasks (CDTs). Again, the stimulus set (oriented coloured arrows) was identical across tasks, this time the

tasks varied systematically with respect to judgement type (orientation alone, colour alone, both orientation & colour together), allowing me to establish them as potential boundary conditions. All participants completed assessment versions of the three CDT tasks and a simple digit span task both before and after training. Each of the assessment tasks had an almost identical training counterpart, one group trained on the orientation-CDT, a second group trained on the colour-CDT task, a third group trained on the orientation & colour-CDT, and finally a fourth group trained on the digit span task (control). Furthermore, both the number of stimuli to-be-remembered and the accuracy with which they were required to be remembered were also varied within each task. This allowed me to go a step further and ask whether the training affected one's ability to retain more items, the quality of those items, or both, and whether these aspects were transferable. Finally, half of the assessment task trials contained a retro-cue indicating the location of the to-be-tested stimulus. This allowed me to ask whether directing attention to items in memory before the response affected the size of the transfer effect; thus, providing a further clue about the cognitive mechanisms targeted by the training. Again, this general design allowed me to ask whether training on different parts of the hierarchy produced different transfer patterns and whether these transfer patterns could be predicted using various metrics of task similarity. However, given the rich theoretical backdrop of the CDT paradigm and the within task manipulations, I was also able to ask more theory driven questions.

Finally, to close the thesis I provide a general discussion on the findings across chapters. This contains a brief re-cap of the main findings from each of the chapters and an attempt to integrate them. I discuss both the strengths and weaknesses of the approaches taken within, as well as the implications they hold for future research in the field of cognitive training.

1.7 Key questions

Below are some of the key questions I hope to answer across the thesis:

i. Are unsupervised machine learning algorithms a viable multivariate alternative for representing, analysing, and interpreting cognitive training data?
ii. Does training alter task relationships?
iii. Are there subgroups of participants with different profiles of change following training?

iv. What types of task relationships best predict transfer patterns following training on different tasks within nested feature-based hierarchies?

v. Are transfer patterns following training on different tasks within nested feature-based hierarchies directional with respect to feature complexity?

Chapter 2: Mapping differential responses to cognitive training using machine learning

2.1 Introduction

2.1.1 Working memory training

Working memory (WM) - the ability to retain and manipulate information for brief periods of time in the service of ongoing task demands, is predictive of healthy cognition across the lifespan and closely linked to academic attainment, employability, and wellbeing (Alloway & Alloway, 2010; Baddeley & Hitch, 1974; Cowan et al. 2005; Diamond, 2012; Johnson et al. 2013). Consequently, the prospect of enhancing WM and closely associated cognitive skills such as attention, processing speed, and reasoning via cognitive training has received considerable interest from researchers and commercial enterprises (Diamond, 2012; Green and Bavelier, 2008, Hertzog et al., 2008; Simons et al., 2016). The assumption being that enhancing this general-purpose system will produce wide benefits to other aspects of cognition and learning.

There have been several promising studies suggesting that training on tasks purporting to tap WM ability or related executive functions may have generalisable benefits to other measures of WM and perhaps even beyond (Green & Bavelier, 2008; Au et al., 2014; Klingberg, 2010; Jaeggi et al., 2008; Holmes et al., 2009). For example, following training on sets of computerised WM tasks that involve the memorisation of both visuo-spatial information (remembering positions of objects in grids) and verbal information (remembering phonemes, letters, or digits), several studies reported transfer to other WM measures not included in the original training program, as well as measures of response inhibition, and complex reasoning (Klingberg et al., 2002; Klingberg et al., 2005; Thorell et al., 2009; Holmes et al., 2009; Holmes et al., 2009). Likewise, others reported transfer to other measures of WM, complex reasoning, and even intelligence following training on the n-back task, a particularly demanding WM measure involving the continual updating of serial information (Dahlin et al., 2008; Li et al., 2008; Jaeggi et al., 2008).

However, many of these findings have failed to replicate and have since been the subject of much scrutiny (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Sala & Gobet, 2019; Simons et al., 2016; Soveri et al., 2017; Von Bastian & Oberauer, 2014). Most of the abovementioned studies trained on a range of tasks making it hard to pin down specific mechanisms of transfer and tease apart task specific from task general effects (as is the case in this study). Moreover, many early findings and interpretations have since been

attributed to the methodological shortcomings such as studies being underpowered, failing to correct for multiple comparisons, or lacking an appropriate control group (Melby-Lervag et al., 2016; Sala & Gobet, 2019; Simons et al., 2016; Soveri et al., 2017).

Consequently, recent experimental WM training studies have implemented more rigorous and tightly controlled designs. Norris et al (2019) compared four groups (Digit-Span training, Circle-Span training, Colour Change Detection training, and a passive control) before and after training on a set of Short-Term Memory (STM) tasks, which are sometimes considered as WM tasks by other researchers and often included in WM training programmes. They found no evidence for cross paradigm transfer effects for any of the groups or within paradigm effects for either the digit or circle-span training groups, even when the assessment task differed only by a single stimulus-feature (e.g. visual vs auditory digit span; visual digit span vs visual letter span). However, they did find strong evidence for within paradigm transfer effects for the Colour-CDT training group to the untrained Orientation-CDT task, which partly motivated the study presented in Chapter 3. These findings provide further evidence for the specificity of transfer effects but also show that this specificity varies between paradigms. Norris et al take these findings as support for the cognitive routine framework (Gathercole et al., 2019). They argue the cognitive routines recruited to perform simple span tasks are already well established by adulthood through commonly encountered activities that also require them and as such these tasks do not require new ones to be performed. Moreover, when people do develop new routines to enhance performance after extensive practice, these tend to be semantically tied to the specifics of the stimuli and thus do not transfer to other span tasks when the stimuli specifications are changed. On the other hand, the Colour-CDT task demands are relatively novel and therefore require the development of a new routine that is not specific to certain stimuli and thus transfers to the Orientation-CDT tasks.

Holmes et al (2019) examined the limits of transfer within and between complex span and n-back training paradigms. Specifically, they compared an n-back training group, a complex span training group, and a passive control group before and after training across 4 assessment tasks: visuo-spatial n-back, verbal n-back, visuo-spatial complex span, and verbal complex span. Crucially, the stimuli were closely matched across the training task paradigms and the assessment task paradigms. Whilst the stimulus category varied between training and assessments, everything else was held constant. This allowed them to parse the extent to which transfer in this study was constrained by training paradigm and/or stimulus material.

Using a Bayesian approach, they found no evidence for cross-paradigm transfer and mixed evidence for transfer from n-back to an untrained visuo-spatial variant. Again, these findings speak to the specificity of training effects both within and between paradigms but, as with previous research (Soveri et al., 2017), indicate that stimulus category is not a boundary condition within the n-back paradigm. Holmes et al (2019) also cite the cognitive routine framework (Gathercole et al., 2019) as a potential explanation but argue that they cannot rule out the enhancement of working memory processes specifically tied to the n-back (e.g. updating) but not to the complex span, as an alternative explanation.

As emphasised in the general introduction, and in line with the rest of the literature, evidence for improvements on tasks similar to those practised (near transfer) following WM training is plentiful; in contrast, evidence for improvements on more distant tasks (far transfer) following WM training is limited (Gathercole et al, 2019; Green and Bavelier, 2008; Hertzog et al., 2008; Melby-Lervag & Hulme, 2016). That is, typical WM training protocols employed to date induce more specific changes to cognition than was previously anticipated and hoped for. Despite an increasing recognition and consensus regarding the specificity of training, evidence is still mixed and many of the conditions required for transfer remain unclear (Gathercole et al., 2019; Melby-Lervag & Hulme, 2016; Sala & Gobet, 2019; Smid et al., 2020). Moving beyond our current understanding of transfer, will require novel lines of enquiry and methodological approaches.

2.1.2 Individual differences and multivariate analysis

As previously mentioned in the General Introduction, the role of individual differences in the size of training effects is receiving increasing attention from researchers. Understanding prior factors that predict transfer effects may help explain many inconsistencies concerning the effectiveness of cognitive training; it could also help tailor training regimes towards those most responsive. Thus far, studies examining individual differences in training paradigms are relatively rare but steadily growing in number. The majority have explored the impact of known pre-training individual differences, such as age (Schmiedek et al. 2010; Borella et al. 2014), baseline cognitive performance (Guye et al., 2017; Bürki et al. 2014; Zinke et al. 2014), and cognition-related beliefs (e.g. malleability of intelligence; Jaeggi, 2014). They provide evidence that some pre-training individual differences may explain variability in training effects.

The majority of these studies have used univariate analytical techniques (e.g. Zinke et al. 2014; Jaeggi, 2014). That is, taking single tasks and testing whether performance on them changes significantly following training, and whether this is moderated by a known individual difference factor. A principal challenge to this approach is task impurity - the extent to which any given task measures an intended construct – because this makes it difficult to identify which mechanisms are being trained (this is intimately related to the task similarity problem highlighted in Chapters 1, 3, and 4; Barnett & Ceci, 2002; Meyer et al., 2001; Hasson, Chen & Honey, 2015; Miyake et al., 2000; Burgess, 2004; Taatgen, 2013; Smid et al., 2020). For example, both n-back and complex span tasks purportedly measure 'WM capacity', but training effects on these tasks do not consistently transfer to one another (Li et al., 2008; Harrison et al., 2013; Holmes et al., 2019). Similarly, both letter span and word span tasks purportedly measure 'verbal short-term memory', but training effects on letter span do not always transfer to word span (Ericcson et al., 1980). In short, the labels assigned to tasks do not always correspond well to the underlying processes taxed by the assessment, or those enhanced via practice. Comparing individual tasks before and after training does not overcome this challenge, because changes on individual measures could stem from changes in multiple different underlying processes (Karbach et al.,2017; Smid et al., 2020; Protzko, 2017; Taatgen, 2013). As a result, a number of researchers are now beginning to explore the potential value of multivariate approaches to considering changes that occur following cognitive training.

One such approach is Structural Equation Modelling (SEM), in which cognitive abilities are represented by latent constructs (Schmiedek et al., 2010; Karbach et al., 2017). Schmiedek and colleagues conducted a large training study, in which they used Latent Score Change Modelling (LSCM, a form of SEM) and found transfer effects to be detectable at a latent level. As they note, it is possible to observe significant changes at the latent level despite non-significant changes at a task-specific level and vice versa. This is presumably because latent constructs may change substantially, but their contribution to any single task in the battery could be relatively small. Conversely, we might observe highly specific practice effects particular to a given paradigm or stimulus set (e.g. letters or digits) that do not stem from changes to any broader underlying latent construct. SEM can be a powerful tool for looking at individual differences because it accounts for measurement error in observed variables and thus provides a good way of establishing stable individual differences (Hamaker et al., 2015). This has enabled some researchers to investigate individual

differences by including separate predictors for the estimated change variable in their models (e.g. Lövdén et al. 2012; Karbach et al., 2017; Guye et al., 2017; Bürki et al. 2014).

Although promising, as with any method, SEM is not without its drawbacks. Confirmatory Factor Analytical (CFA) approaches require researchers to make subjective choices (albeit based on theory) about the structure of underlying components from the many possible configurations, at differing levels of granularity. Furthermore, establishing training effects is particularly challenging because the nature of the underlying constructs, their interrelationships, or their task loadings may have changed substantially as a function of the training. Investigators are faced with a dilemma: they can fit the same model both before and after training, allowing for a meaningful comparison of model parameters but ignoring the fact that this model may no longer be the most appropriate. Alternatively, they can fit the best model separately before and after training, which would allow for the best representation of the underlying components but render direct comparisons less meaningful.

2.1.3 Machine learning approach

Machine learning provides an alternative to modelling task relationships. Unsupervised learning algorithms hold the same advantage as other data-driven methods such as Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA), in that they allow researchers to explore task relationships without requiring subjective judgements to be made about their nature a priori. Machine learning algorithms also lend themselves well to non-linearities in multidimensional data, allowing them to capture more nuanced task relationships compared with commonly used linear methods (general linear regression, factor analysis, PCA etc.). Some algorithms cluster participants in a competitive manner, rather than clustering tasks at the whole-group level (as would be the case for PCA or EFA). These may be particularly useful when we suspect there could be large individual differences – in the context of training, resulting in differing profiles of change following an intervention. Iterative clustering techniques provide a data-driven way of subgrouping participants and thereby reveal different profiles of performance. Potentially, they could enable researchers to explore individual differences in training in a different way – rather than testing whether gains in training are predicted by known factors (e.g. age, baseline ability), they allow researchers to identify individual differences in the profile of the training response itself. Despite these potential benefits, there appears to be little attempt to use machine learning to understand transfer effects following cognitive training. This paper aims to explore the utility of combining two relatively simple machine learning techniques: namely, Self-Organising

Maps (SOMs) and K-Means clustering, to explore task relationships and how these might be altered by training in two large datasets.

First proposed by Kohonen (1989), SOMs belong to a family of artificial neural networks and provide a way of organising multidimensional data into a lower dimensional space, represented as a topographical distribution. An unsupervised learning algorithm projects the original data from a multidimensional input space onto a two-dimensional grid of nodes called a map. Each node corresponds to a node-weight vector with the same dimensionality as the number of input variables, thereby producing an inter-variable representational space, wherein the geometric distance between nodes corresponds to the degree of similarity in the input data associated with them (Kohonen, 2014). This enables key inter-variable relationships existing in multidimensional space to be identified and accentuated. Moreover, this allows the researcher to explore the overlap in this representational space between tasks and how, or whether, it changes as a result of the training. Once established, SOMs may be used to generate quantitative predictions about training effects in unseen data, something currently underutilised in cognitive training research (Barnett & Ceci, 2002; Taatgen, 2013; Gathercole et al., 2019).

Subsequently, a K-means clustering algorithm can be used to identify relatively homogenous subgroups (i.e. ‘clusters’) within the multidimensional node-weight vector space produced by the SOM algorithm. This allows for the exploration of individual differences in task relationships and makes use of information that would otherwise be lost. Identifying data-driven subgroups with distinct cognitive profiles could prove a valuable way of understanding different trajectories in cognitive change.

2.1.4 The present study

Due to shortcomings in explaining and pinpointing the efficacy of WM training, attention has been drawn to the potential benefits of further exploring both individual differences and the use of multivariate approaches toward investigating the effects of WM training and cognitive training more generally. As such, the present study explores the combined use of two unsupervised machine learning algorithms (SOMs and K means clustering) as alternatives for looking at individual differences in responses to training. More specifically, I used these algorithms to model multivariate data across four popular WM tasks in a large composite dataset and to identify subgroups with respect to performance profiles. I then used these models and performance profiles to address the following three questions:

1. Do the SOM models represent the samples well?
2. Do SOM models generalise across samples?
3. Are there subgroups with different profiles of change following training?

2.2 Materials and methods

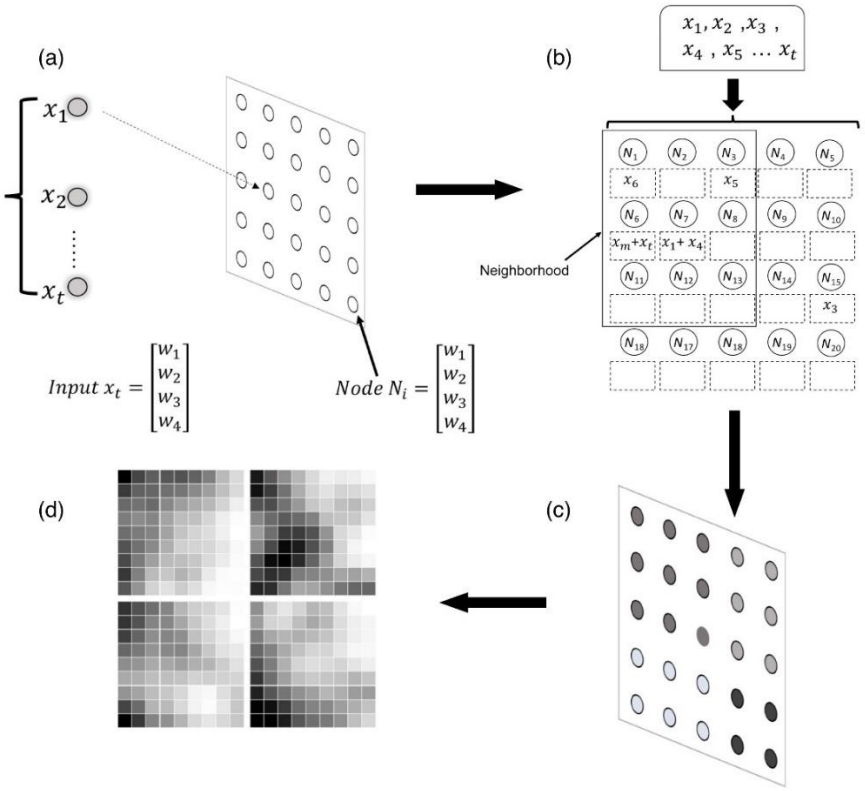
This section contains a brief description of the Self-Organising Map (SOM) algorithm and its generic implementation, followed by a stepwise account of the analyses performed on two datasets containing the same set of tasks.

2.2.1 SOM algorithm

SOMs were trained using the neural network toolbox in Matlab (MathWorks, 2017a). SOMs consist of a predefined number of nodes laid out on a two-dimensional grid plane. Each node corresponds to a weight vector with the same dimensionality as the input data. The node weight vectors were initialised using linear combinations of the first two principal components of the input data. SOMs were then trained using a batch implementation (see Figure 2.1 for a graphical overview), in which each node $\boldsymbol{i}$ is associated with a model $\boldsymbol{m_i}$ and a 'buffer memory'. One cycle of the batch algorithm can be broken down into the following: Each input vector, in this case a single child's performance profile across the four assessment tasks, $\boldsymbol{x(t)}$ is mapped onto the node with which it shares the least Euclidean distance at time $\boldsymbol{t}$. This node is known as its Best Matching Unit. Each buffer sums the values of all input vectors $\boldsymbol{x(t)}$ in the neighbourhood set belonging to node $\boldsymbol{i}$ and divides this by the total number of these input vectors to derive a mean value. All $\boldsymbol{m_i}$ are then updated concurrently according to these values. In this way, neighbouring nodes become more similar to one another. This cycle is repeated, clearing all the buffers on each cycle and distributing new copies of the input vectors into them. The neighbourhood size ($\boldsymbol{ND}$) decreases as a function of $\boldsymbol{t}$ over $\boldsymbol{n}$ steps in an 'ordering' phase, from the initial neighbourhood size ($\boldsymbol{INS}$) down to 1 (Equation 1.). In the 'fine tuning' phase the neighbourhood size is fixed at <1, meaning that the node weights are updated according only to the input vectors for which they are the Best Matching Unit. This node adjustment process is the mechanism by which the SOM learns about the input data.

Equation 1. $$\boldsymbol{ND = 1 + \left[(INS) * \left(1 - \left(\frac{t}{n}\right)\right)\right]}$$

Figure 2.1 Illustration of SOM batch training steps to update node weights using given dataset.



Note. (a) Each input vector $x_{(t)}$ is mapped onto its Best Matching Unit. (b) All input vectors in each node are summed and used to update its Best Matching Unit and neighborhood, which shrinks with time. (c) When training completes, SOM has preserved the topological information of the input data. Data with similar inter-variable relationships are assigned to closer Best Matching Units. (d) Visualisation of individual node weights (ω_n), namely the component planes, when input vector contains 4 variables.

2.2.2 Assessment tasks

Four span tasks from the Automated Working Memory Assessment battery (AWMA; Alloway et al., 2008) were used in the current analysis. In Forward Digit Recall, participants hear a sequence of numbers and are required to repeat them back out loud in the same order in which they were presented. In Backward Digit Recall, participants hear a sequence of numbers and are required to repeat them back out loud in the reverse of the presentation order. These tasks are purported to measure verbal short-term memory and working memory, respectively. In Dot Matrix, participants see a sequence of dots in a 4x4 matrix and are required to recall the order and position of the dots by pointing to a blank 4x4 response matrix. In Mr.X, participants are presented with sequences of two cartoon characters placed next to one another, both of which are holding a ball in one of their two outstretched arms, and the one on the right is rotated to varying degrees on each presentation. For each pair of Mr. X's participants are required to make a same-different judgement with regards to whether they are holding the ball in the same hand or not, whilst retaining the spatial information as to

where the ball held by the right-hand Mr. X resides. They are then required to recall the previously retained spatial locations in the correct order by pointing to one of six locations represented by dots in a circle. These tasks purport to measure visuospatial short-term memory and working memory, respectively. All tasks along with the instructions are computerised and practice trials were completed on each to help ensure comprehension.

2.2.3 Participants

There were three relatively large datasets used in this analysis. All datasets consisted of age-standardised data (mean=100, standard deviation=15) from the four AWMA tasks. Each of these is described in the following sections, and summary scores are described in Table 2.1.

Centre for Attention, Learning and Memory (CALM)

The first dataset comprised of data collected from 526 participants (M= 9.16 years, Range: 5.16-17.91 years, SD= 2.16 years; 171 girls) who had completed assessments as part of the CALM study. This is a study of children referred based on ongoing problems in attention, learning and memory. Children visit the MRC Cognition and Brain Sciences Unit and undergo a wide battery of cognitive and behavioural assessments, which includes the four tasks described above.

Attention and Cognition in Education (ACE)

This sample was collected for a study investigating the neural, cognitive, and environmental markers of risk and resilience in children. Ninety typically developing children who attend mainstream schools in the UK (M= 9.42 years, Range: 6.91-12.58 years, SD=1.49 years; 45 girls) and their families were invited to the MRC Cognition and Brain Sciences Unit in Cambridge for a comprehensive cognitive assessment, which included the four tasks described above.

In later analyses, the data from the two abovementioned studies was combined for greater statistical power and larger individual variability in task profiles, which is desirable for a “baseline” dataset.

Table 2.1. Summary statistics of task performance in the respective datasets.

Dataset			Forward Digit Span	Dot Matrix	Backward Digit Span	Mr.X
CALM (N=526)		Mean	91.93	91.55	90.80	97.45
		SD	15.98	15.17	13.44	14.86
ACE (N=90)		Mean	104.14	103.71	103.65	105.58
		SD	13.38	14.97	14.77	15.60
CALM+ACE (N=616)		Mean	93.86	93.46	92.81	98.64
		SD	15.84	15.41	14.00	15.26
Pre-Training	Adaptive (N=179)	Mean	93.95	90.78	85.58	89.73
		SD	15.58	16.12	14.90	16.16
	Non-adaptive (N=70)	Mean	90.93	94.29	84.84	91.34
		SD	16.27	16.93	13.04	15.88
Post-Training	Adaptive (N=179)	Mean	100.54	110.56	101.01	102.53
		SD	17.64	19.59	15.43	18.68
	Non-adaptive (N=70)	Mean	91.94	103.51	96.32	100.83
		SD	18.01	18.89	17.23	20.91

Note. The expected mean and standard deviation of the normative AWMA data is 100.00 and 15.00, respectively.

Combined training studies

This dataset comprised of Pre-training and Post-training data collected from 179 participants (M=9.00 years, Range:7.08-11.50 years, SD=1.06 years; 45 girls), combined over several independent training studies (Dunning, et al., 2013; Holmes et al., 2009; Holmes et al., 2010; Holmes et al, 2015). Inclusion criteria varied across studies, such as low WM score on standard tests (Dunning et al, 2013; Holmes et al., 2009), low language abilities (Holmes et al., 2015) or ADHD diagnosis (Holmes et al, 2010). All children participated in the standard Cogmed RM program (see: Klingberg et al, 2005, for a detailed description of the training tasks), which involved 20–25 sessions of adaptive training on temporary storage and manipulation of sequential visuospatial or verbal information, or both. Detailed methods regarding the training program have been reported previously (Klingberg et al, 2005). The same four AWMA tasks were administered before and after training as measures of transfer, leading to a total of 8 variables for this dataset. Participants showed significant improvements on all four tasks in the Post-training assessment ($p < .001$) compared to their baseline assessments, most notably on Dot Matrix and Backwards Digit Span (Cohen's d: Forward Digit = 0.395, Dot Matrix = 1.103, Backward Digit = 1.017, Mr. X = 0.732). Summaries are also included for a combined control group who were given a non-adaptive version of the Cogmed training (M = 9.02 years, Range: 7.50-10.50 years, SD = 0.72 years; 29 girls). The corresponding repeated measures Analysis of Variance (ANOVA) established a significant treatment by time interaction for Forward Digit, $F (1, 69) = 4.58$, $p <.05$; Dot Matrix, F (1,

69) = 27.36, p <.001; Backward Digit, F (1, 69) = 9.62, p < .01; and Mr. X, F (1, 69) = 4.59, p <.05. In all cases, the improvements were significantly greater for the training group than the control group. Furthermore, simple main effect analysis showed that the performances of control group on all tasks were significantly better at post-training than pre-training (p < .001), except for Forward Digit (p = .48).

2.4 Analysis pipeline

2.4.1 Training SOMs

The SOM learning algorithm and model require the selection of several parameters, including the number of map nodes, initial neighbourhood size, the ordering phase length, and fine-tuning phase length. These hold important theoretical, computational, and statistical implications. However, according to Kohonen (2014), there are no standard mathematical definitions to inform the selection of such parameters. Instead, Kohonen covers some key concepts and provides suggestions based on experience. A detailed discussion of this topic is beyond the scope of this thesis. However, for a more detailed explanation of our selection process and an overview of the results, see Appendix A at the end of the thesis. In short, parameters were selected with the aim that the SOM model would represent the training sample well, whilst still maintaining generalisability to the wider population. Three SOMs were trained: 1) a SOM trained on the combined CALM and ACE dataset (CALM/ACE); 2) a SOM trained on the Pre-training dataset; and 3) a SOM trained on the Post-training dataset. The relatively large sample size of CALM/ACE (616 participants) provided a good baseline dataset for learning about the overlap between the different tasks, and the possible cognitive profiles that exist. The smaller training datasets were used to investigate questions about training effects.

2.4.2 Do the SOM models represent the samples well?

The first step after fitting a model is to test its validity. A cross-validation procedure was applied to test the null hypothesis that the SOM does not estimate unseen data above chance levels. Specifically, this involved randomly removing 20% of the CALM/ACE data (i.e. approximately 120 participants), then using the remaining 80% to fit a SOM, which was used to predict the reserved data. The prediction was made with a technique called K-Nearest Neighbours (KNN; Altman, 1992), in which the value of the to-be-predicted variable is decided by the values of the 3 closest SOM nodes in terms of Euclidian distance with respect to the vector containing the other-unseen variables. For example, if Forward Digit is the

target variable, a subject's scores on the other 3 tasks will be fed to the algorithm to find the 3 nearest SOM nodes. Then the values of the three nodes on Forward Digit are pooled and weighted based on distance (the closest node has the highest weight) to calculate the participant's predicted score. The mean absolute difference between the predicted scores and true scores of the unseen sample was used as the measure of prediction error.

To better evaluate the average model performance, the cross-validation process was repeated 1,000 times to derive distributions of the mean prediction errors. The distributions for chance level were achieved by randomly shuffling the order of the predicted scores, then subtracting the true scores to obtain a null mean absolute difference.

For each iteration of the 1,000 cross-validations, the shuffling was also repeated 100 times to create a null distribution containing 100,000 values of prediction errors. Finally, the mean prediction errors for all variables were compared to the corresponding null distributions to derive p-values by calculating the proportion of the null distribution greater than the mean prediction error.

2.4.3 Do SOM models generalise across samples?

I was also interested in whether the representativeness of the SOM extended to other samples. To test this, a SOM was trained on the entire CALM/ACE dataset and used to predict task scores in the Pre- and Post-training datasets. The CALM/ACE sample is much larger in size and includes a wide range of ability levels. This means that a model based on these data is more likely to generalise well to other datasets. Chance level distributions were generated for Pre- and Post-training samples similarly to the last step by shuffling the order of predicted scores 100,000 times. Again, true prediction errors were compared to derive p-values.

An alternative way to address this question is to compare prediction errors for the CALM/ACE sample and the Pre- and Post-training samples respectively. If the SOM model represents the training study data as well as it does the CALM/ACE sample, then prediction errors should not differ from each other. For this purpose, I repeated the same cross-validation procedures 1,000 times but randomly removed 179 participants from the CALM/ACE sample each time to keep the number consistent with the size of the Pre- and Post-training data. The remaining CALM/ACE data points were used to train a SOM and make predictions for the removed CALM/ACE participants, as well as the Pre and Post training samples respectively. A permutation test followed to test for significance of the

difference in prediction errors between CALM/ACE and the training data (i.e. CALM/ACE vs. Pre; CALM/ACE vs. Post). I also used permutation tests to compare the prediction errors of the pre-training and post-training data respective to one another.

2.4.4 Does training alter the relationships between tasks?

Here I ask this question in two ways. Each model node is an instance of a multivariate task relationship that exists in the data used to train the SOM. If these SOM maps have less predictive power when used to estimate new data points this means that different multivariate task relationships exist in that dataset, which are not well accounted for by the model. This is the first way of testing whether the training has changed task relationships.

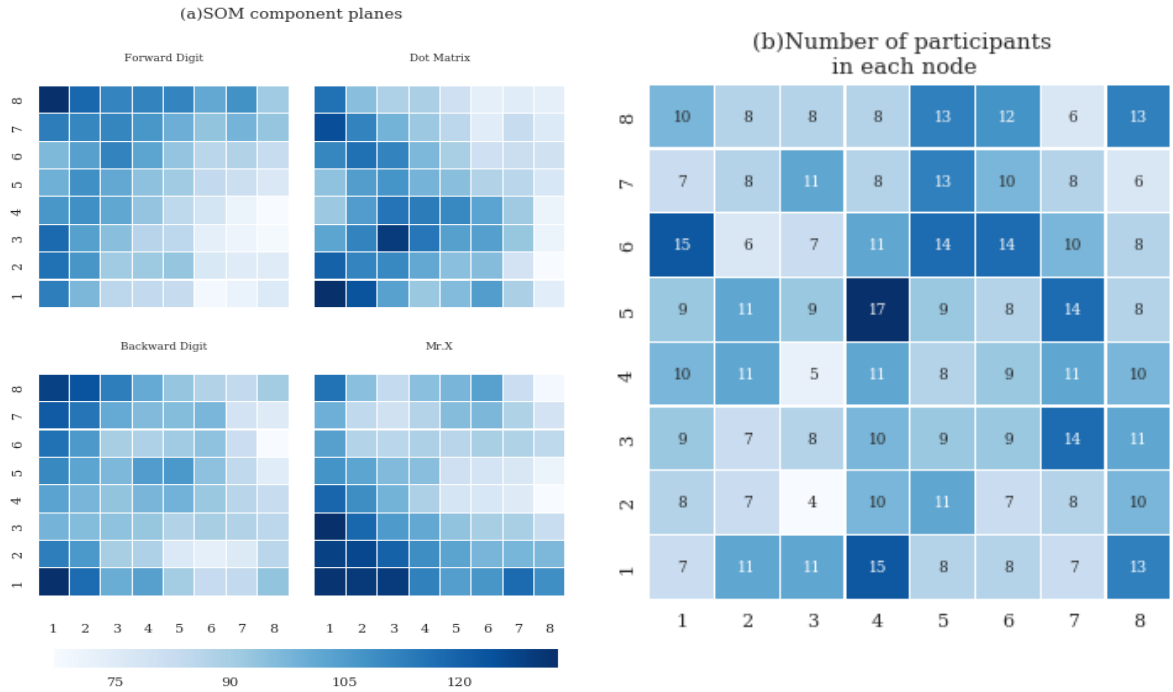
The second way of addressing this question was by comparing the SOMs trained on the Pre- and Post-training datasets directly. If two tasks tap into similar cognitive processes, their model representations should overlap; if training alters task relationships, then the model representations of tasks in the SOM fit to Pre-training data should be substantively different to those in the SOM fit to Post-training data.

To access task similarities as represented by SOM, elements of SOM node-weight vectors can be extracted individually (e.g. the 1st element of all node vectors) to form a 'component plane'. Each plane corresponds to a representation of a task. The pairwise correlation coefficients between component planes can be derived and serve as multivariate activity patterns, which is useful for quantitative analysis. If two tasks tap into similar cognitive processes, their activity patterns ought to overlap (e.g. in Figure 2.2 the Forward Digit and Backward Digit which both involve auditory information, share more topological similarity). By extracting the correlations between the same pair of tasks before and after training, we could then make a direct comparison of how their relationships had changed as a result of the training. To compute the relationship, component planes associated with each pair of tasks were compared using Pearson's correlation coefficient. Then, the similarity values are assembled into a 4x4 matrix.

Once the similarity matrices for Pre- and Post-training were computed, the same pairs of tasks between times of Pre- and Post-training were compared to identify any significant differences in correlation coefficients. I chose to bootstrap the node weight elements associated with the two tasks and computed the correlation coefficients before subtracting one from another (Post-training pairwise correlation – Pre-training pairwise correlation). By repeating this procedure 10,000 times, a distribution of the difference between correlations

was obtained. If zero fell within the bottom or top 5% of the distribution, the null hypothesis that the two correlation coefficients are not different was rejected, with a false positive rate of $\alpha = 0.05$. This analysis was conducted for all pairs of tasks.

Figure 2.2 Overview of SOM model trained on CALM/ACE sample



Note. (a) Visualisation of node weights of the SOM (component planes) separated by each task. (b) Number of participants allocated to each node.

2.4.5 Are there subgroups with different profiles of change following training?

K-means clustering provides a data-driven method for identifying k relatively homogenous subgroups within the SOM node-weight vector space by minimizing the distance between data points and the centroids of each cluster (MacQueen, 1967). Although there is no clear theoretical rationale for the choice of number of clusters, in Appendix A multiple cluster solutions are included to demonstrate the resulting differences from various choices of k, also included is a Silhouette Analysis of the different clustering solutions as a measure of the clustering quality.

First, subgroups were identified within the SOM fit to the CALM/ACE data, by applying k-means clustering to the node weights. Once the nodes were grouped based on similarity, participants were allocated to the cluster to which their Best Matching Unit belonged. This provided clusters of children based on the nodes to which they were assigned in the original mapping. This process was repeated 1,000 times, with the map being retrained on every iteration and the k-means clustering being recalculated, to check that the clusters

were robust. Participants in the training datasets were also allocated to these identified clusters in the same manner (i.e. based on closest Euclidean distance) at both Pre- and Post-training, separately. Profiles of subgroups were characterised by calculating their respective means and standard errors on each of the tasks and compared between groups to identify the ways in which they differed. In the case of the cognitive training datasets, the children who changed subgroup following the training were also contrasted against one another. This was done by calculating their gain scores (Post- minus Pre-training) on each task and used to test how different gain scores are associated with changes in subgroup membership.

Finally, I tested whether these clusters were predicted by another measure that was not included in the SOM training or clustering, namely matrix reasoning scores. Importantly this is not a baseline outcome assessment nor in the training regime. Scores on a matrix reasoning task taken from Wechsler's Abbreviated Scale of Intelligence (WASI; Wechsler, 2011) were available for 158 participants in the training sample. Matrix reasoning is considered a measure of general fluid intelligence (Gf), which refers to the ability to reason and solve novel problems. Gf is a critical factor for success in a wide variety of cognitive tasks and the capacity to learn in general (Gary & Thompson, 2004). I explored whether performance on the WASI matrix reasoning task assessed prior to training was predictive of change of subgroup membership.

2.4.6 Analysis summary

The above pipeline describes the stepwise analyses. 1) SOMs were used to model task relationships. Cross-validation was used to test the reliability of the model trained on the large CALM/ACE sample and also whether it was representative of the Pre- and Post-training samples. 2) SOM models representing the Pre- and Post-training samples were compared directly by training new SOMs with the two samples and then comparing them through a representational dissimilarity analysis that examined how task relationships changed following training. 3) K-means clustering was used to identify relatively homogeneous cognitive profiles in the CALM/ACE sample as represented by the SOM model. 4) Participants in the training dataset were subsequently mapped to these subgroups to investigate the changes in these profiles as a function of training. 5) I tested whether fluid intelligence predicted the change in profiles following training.

2.5 Results

A 64 node (8x8) SOM with an initial neighbourhood size of 2 was trained over 10 ordering phase steps and 2 fine tuning phase steps using the CALM/ACE data (quantisation error = 9.72); quantisation error is defined as the mean absolute distance between the input vectors (i.e. training data) and their corresponding Best Matching Units. The rationale behind the selection of these parameters, alongside different solutions with different parameters, is included in Appendix A. Figure 2.2 shows how the SOM represents the four tasks as well as the number of participants allocated to each node.

2.5.1 Do the SOM models represent the samples well?

First, the model performance of the SOM trained on the CALM/ACE data was cross-validated using permutation testing. The SOM proved capable of predicting unseen CALM/ACE data significantly better than chance for all four task variables (Table 2.2).

Table 2.2. SOM prediction errors for the CALM/ACE sample, Pre- and Post-training sample respectively

		Forward Digit	Dot Matrix	Backward Digit	Mr.X
CALM/ACE	Prediction error	11.68	11.57	9.41	11.91
	p	<.001***	<.001***	<.001***	<.001***
Pre-training	Prediction error	13.41	12.24	11.83	14.13
	p	<.001***	<.001***	<.001***	<.001***
Post-training	Prediction error	14.25	17.67	12.12	14.71
	p	<.001***	<.001***	<.001***	<.001***

Note. Standard scores for the prediction error were defined as mean absolute difference between the predicted scores and true scores. *P*-values were derived from comparing the prediction errors against the corresponding chance level distributions. The chance levels were achieved by randomly shuffling the order of the predicted scores, then subtracting the true scores for 100 times within each cross-validation literation, to obtain a null distribution of mean absolute difference. Asterisks denote statistical significance at *p< .05, **p < .01 or ***p< .001.

2.5.2 Do the SOM models generalise across samples?

Next, a SOM trained on the entire CALM/ACE sample was used to test how well it represents the Pre- and Post-training datasets, again using the same method. Again, the model predicted unseen data from other samples better than chance on all tasks.

The CALM/ACE prediction errors were also compared directly with the Pre- and Post-training data (Table 2.3). Predicting the remaining CALM/ACE sample was more accurate than predicting the Pre- or Post-training samples. A direct comparison of the Pre- and Post-training prediction accuracies revealed comparable prediction accuracies on all tasks except on the Dot Matrix task, wherein the prediction accuracy dropped significantly for the post training sample.

Table 2.3. Direct comparisons between SOM prediction errors across samples

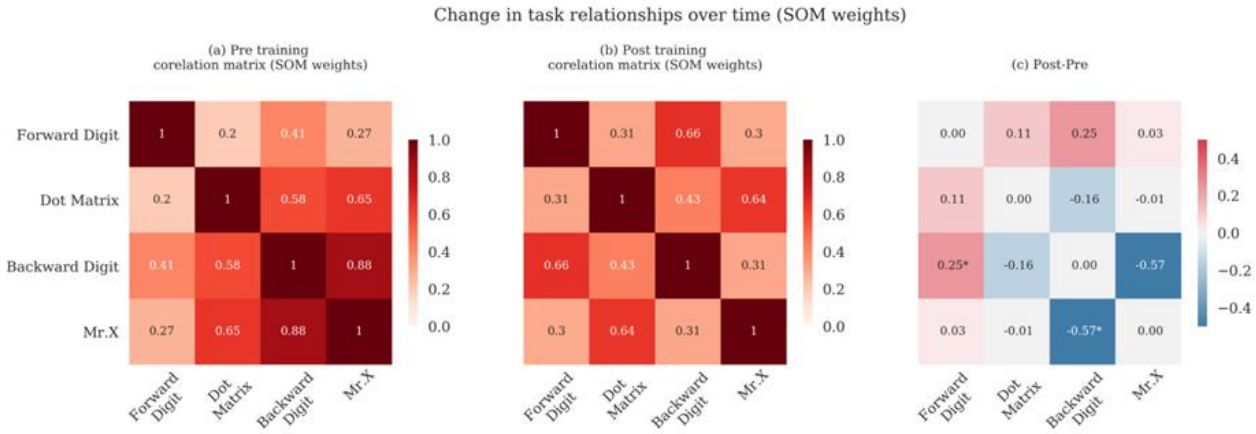
		Forward Digit	Dot Matrix	Backward Digit	Mr.X
CALM/ACE vs Pre-training	Prediction error difference	1.73	0.66	2.42	2.21
	p	<.001***	0.177	<.001***	<.001***
CALM/ACE vs Post-training	Prediction error difference	2.57	6.09	2.70	2.80
	p	<.001***	<.001***	<.001***	<.001***
Pre vs. Post-Training	Prediction error difference	0.83	5.43	0.28	0.59
	p	0.201	<.001***	0.373	0.296

Note. *p*-values were derived from permutation tests. *p< .05, **p < .01 or ***p< .001.

2.4.3 Does training alter the relationships between tasks?

New SOMs trained on Pre- and Post-training data respectively were compared to examine changes in task relationships as a function of training. Pairwise correlation coefficients were computed from the SOMs component planes representing tasks and assembled into similarity matrices. Figure 2.3a and 2.3b depict the Pre- and Post-training matrices. Pairwise comparisons were conducted before and after the training to understand whether there were specific alterations between any of the task relationships. Permutation testing indicated a significant difference in the Backward Digit-MR.X pair (p< .01) and in the Forward Digit- Backward Digit pair (p< .05). In other words, the way tasks were represented in the SOM weights changed following training, with some becoming more similar and others more dissimilar (readers are referred to Figure A.10 and the section titled ‘Comparison with the control group’ in Appendix A. for the same analysis in the control data).

Figure 2.3 Pairwise task relationships derived from SOM weights before and after training and the difference over time.



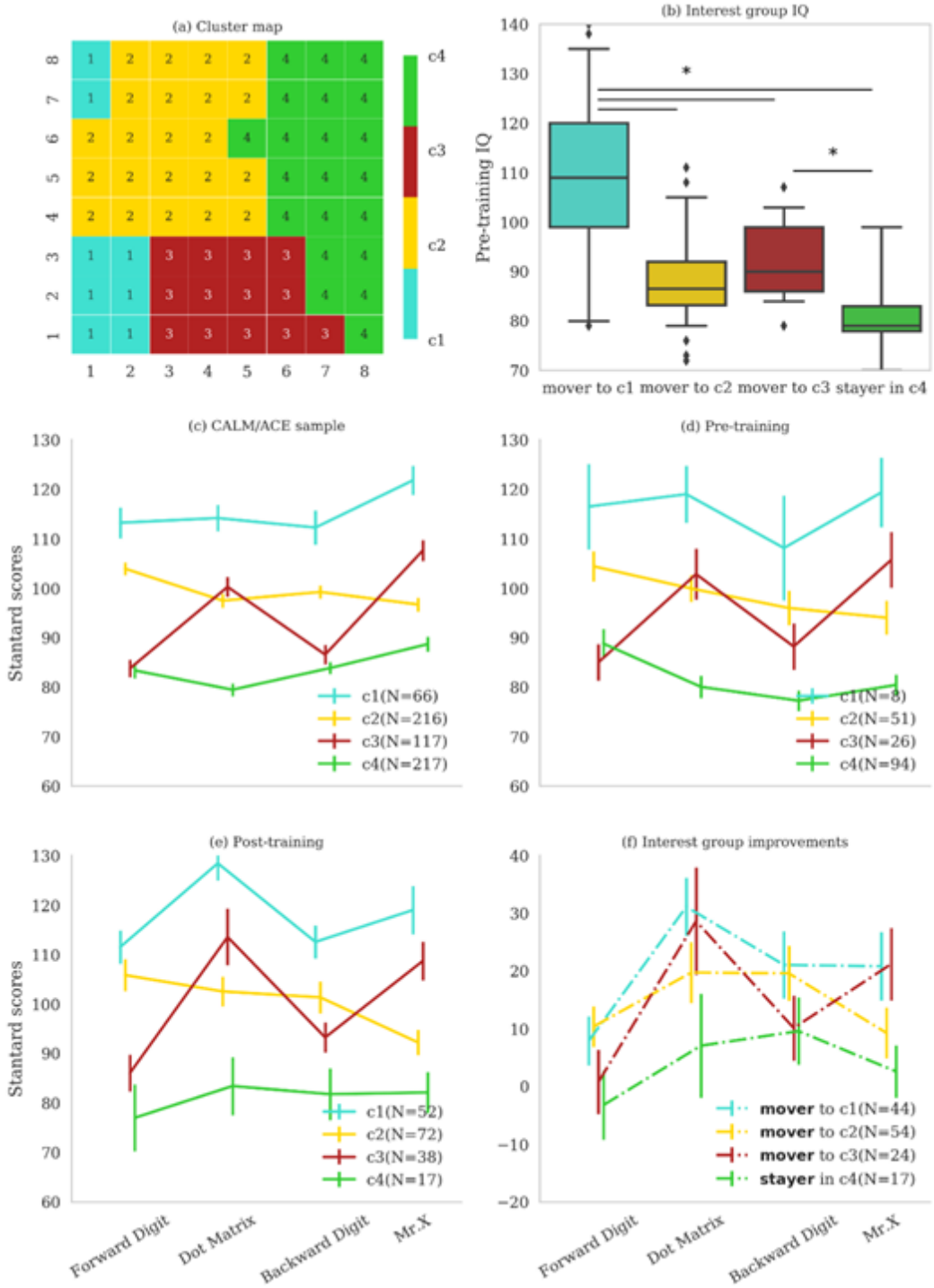
Note. Larger value indicates more similarity between the two tasks. (a) Task relationships for Pre training sample. (b) Task relationships for Post training sample. (c) Difference in similarity between Pre and Post training (Post – Pre). The Backward Digit- Mr. X pair showed significant change after training (p <.01), as did the Forward Digit-Backward Digit pair (p<.05).

2.4.4 Are there subgroups with different profiles of change following training?

K-means clustering was applied to the node-weight vector space pertaining to the SOM trained on the CALM/ACE data using K = 4 (see the Appendix A for robustness of clustering quality across different Ks). The resulting partition of the SOMs nodes can be seen in Figure 2.4a. Participants were allocated to the cluster to which their Best Matching Units belonged. Profiles of subgroups were characterised by calculating the respective means and standard errors on all four tasks in the CALM/ACE sample (Figure 2.4c).

Between-group ANOVAs were conducted for each task and all indicated significant differences (Forward Digit: F (3,612) = 233.17; Dot Matrix: F (3,612) = 244.03; Backward Digit: F (3,612) = 175.59; Mr. X: F (3,612) = 179.41; all p <.001). Results from the post-hoc Tukey's HSD tests showed that all subgroups differed significantly from one another at the .05 level except for group 2 and 3 on Dot Matrix and group 3 and 4 on Forward Digit and Backward Digit. The algorithm identified a subgroup of participants who achieved a high level of performance on all tasks, a subgroup whose scores were at the lower end of the distribution and two subgroups who were in the middle. One of these middle subgroups tended to have average performance on all tasks, whereas the other tended to have average or slightly above average performance on the visual spatial tasks but below average performance on the verbal tasks.

Figure 2.4. Results of K-mean clustering and comparison of subgroup profiles.



Note. (a) SOM nodes were partitioned into 4 clusters. (c, d & e) Comparison of task scores among the subgroups in the CALM/ACE, the Pre-training and the Post-training sample respectively. Error bar indicates 95% confidence interval. (f) Comparison of improvement profiles of the interest groups: participant who moved to the highest-performing group (cluster 1) after training, those who moved to the medium group with verbal-specific gains (cluster 2), those who moved to the medium group with visuospatial-specific gains (cluster 3), and those who stayed in the low-performing group (cluster 4). For clarity, the first three interest groups only included participants who moved to the respective group from the outside, not those that were already there at pre-training, and vice versa for the fourth interest group. (b) WASI matrix reasoning score of the three interest groups. *p<.05.

Participants in the Pre- and Post-training sample were also allocated to one of the four identified subgroups (Figure 2.4d & 2.4e). The profiles of the training study participants in each subgroup were similar to those of the CALM/ACE sample, highlighting the ability of K-means clustering to determine relatively homogenous groups (see Appendix A). Specifically, ANOVAs indicated significant group differences across all measure for Pre and Post-training (Pre-training: Forward Digit: $F(3,175) = 28.58$; Dot Matrix: $F(3,175) = 69.84$; Backward Digit: $F(3,175) = 40.76$; Mr. X: $F(3,175) = 55.67$; all $p < .001$; Post-training: Forward Digit: $F(3,175) = 50.31$; Dot Matrix: $F(3,175) = 58.08$; Backward Digit: $F(3,175) = 33.63$; Mr. X: $F(3,175) = 55.18$; all $p < .001$). Post-hoc Tukey's HSD tests revealed all pair-wise groups were significantly different at the .05 level except for group 2 and 3 on Dot Matrix and group 3 and 4 on Forward Digit in the Pre-training dataset; difference between group 1 and 2 on Forward Digit was marginally non-significant ($p = .06$). For the Post-training, all subgroups were different from one another.

The gain scores (Post-training minus Pre-training scores) of children who moved to a different subgroup following training were calculated and contrasted with one another to capture individual differences in training–related improvement. Four interest groups were identified (separate from but related to the original four clusters): children who moved to the highest performance group (mover to cluster 1), children who moved to group 2, children who moved to group 3 and children who stayed in the lowest performance group (stayer in cluster 4). For clarity, the first three interest groups only included participants who moved to the respective group from the outside, not those that were already there at Pre-training, and vice versa for the fourth interest group. The gain scores of these groups were shown in Figure 2.4f. These groups were significantly different from each other according to ANOVA (Forward Digit: $F(3,122) = 4.94$; Dot Matrix: $F(3,122) = 7.14$; Backward Digit: $F(3,122) = 6.34$; Mr. X: $F(3,122) = 8.57$; all $p < .001$) and post-hoc tests (see Table 2.4 for multiple pairwise comparison results). Overall, movers to 1 had the highest improvement across all measures compared to the other groups. Movers to 2 were characterised by moderate gains globally but benefited less on Dot Matrix and Mr.X relative to children moved to cluster 1. The third group, children moved to cluster 3 had comparable magnitude of gains on Dot Matrix and Mr.X to movers to 1, but significantly less gains on Forward and Backward Digit tasks than movers to 1 or 3. Unsurprisingly, children who stayed in the lowest performance group (cluster 4) had an overall limited benefit from the training.

Table 2.4. Results of multiple comparison between different improvement profiles.

	Movers to C1	Movers to C2	Movers to C3	Stayers in C4
Forward Digit				
Movers to C1	NA	p = 0.88	p = 0 .36	p < 0.05*
Movers to C2		NA	p = 0 .09	p < 0.01**
Movers to C3			NA	p = 0.63
Movers to C4				NA
Dot Matrix				
Movers to C1	NA	p = 0 .08	p = 0.90	p < 0.001***
Movers to C2		NA	p = 0.18	p = 0.06
Movers to C3			NA	p < 0 .01**
Movers to C4				NA
Backwards Digit				
Movers to C1	NA	p = 0.71	p < 0.01**	p < 0.01**
Movers to C2		NA	p < 0.05*	p = 0.06
Movers to C3			NA	p = 0.90
Movers to C4				NA
Mr.X				
Movers to C1	NA	p < 0.01**	p = 0.90	p < 0.001***
Movers to C2		NA	p = 0.05	p = 0.30
Movers to C3			NA	p < 0.01**
Movers to C4				NA

Note. *p< .05, **p < .01 or ***p< .001.

To investigate whether performance on a measure of fluid intelligence (WASI matric reasoning) could predict these individual differences in patterns of improvement, the four interest groups' WASI scores assessed prior to training (see Figure 2.4b) were compared. ANOVA indicated significant WASI score differences among the groups ($F(3,122) = 28.83$, $p < .001$). Post-hoc tests showed that movers to cluster 1 had higher WASI scores ($M = 108.72$, $SD = 17.39$) than movers to cluster 2 ($M = 88.46$, $SD = 8.70$, $p < .001$), movers to cluster 3 ($M = 92.31$, $SD = 7.94$, $p < .001$), and stayers in cluster 4 ($M = 81.76$, $SD = 8.72$, $p < .001$). Movers to 3 also had higher WASI scores compared to stayers in 4 ($p < .05$).

2.6 Discussion

Our understanding of cognitive training hitherto has focused primarily on exploring its impact on single tasks (though with some notable exceptions, e.g. Guye et al., 2017; Karbach et al., 2017; Schmiedek et al., 2010) and treating all participants as a single homogenous group (e.g. Borella et al. 2014; Guye et al., 2017; Bürki et al. 2014; Zinke et al. 2014). The present study used machine learning to show that working memory training alters the relationships between tasks, implying that the cognitive processes recruited for performing those tasks may have changed following training. Furthermore, subgroups with differential responses to training were identified and predictive of fluid intelligence scores.

2.6.1 SOMs accurately represent task relationships

A SOM was fit to a large dataset of children who were assessed on four standardised measures of verbal and visuospatial short-term and working memory. Leave-N-out cross-validation showed that SOMs fitted on these data predicted performance on unseen data for all tasks. These predictions generalised to the cognitive training samples; importantly, however, the model fit and prediction accuracy was reduced significantly following training for the Dot Matrix, the implication of which is discussed subsequently.

2.6.2 Task relationships change following training

Multiple studies have shown that performance on individual tasks improves following training (for reviews, see Hertzog et al. 2008; Melby-Lervag and Redick, 2016; von Bastian and Oberauer 2014). But this provides limited insight into whether or how underlying constructs are being changed, or whether different cognitive processes are recruited following the intervention. One way of investigating this is to test whether relationships between tasks change as a function of training. Following training, there was a large decrease in model prediction accuracy for the Dot Matrix task mirroring substantial improvements in task performance. Lower prediction accuracy following training also suggests that the relationships between Dot Matrix and the other tasks may have been altered. In other words, new task relationships (i.e. multivariate data points) exist in the post-training data that were not learnt or represented in a large sample of children who did not complete the cognitive training. In this case, the training programme contains a lot of exercises similar to the Dot Matrix task (i.e. visuo-spatial serial recall, Klingberg et al., 2005), and thus subjects may show a more task-specific effect rather than a domain general improvement. This would be in line with research demonstrating that transfer tends to be tied to specific task features, with the highest levels of transfer found on assessment tasks that share the greatest number task features with those trained (Gathercole et al., 2019; Soveri et al., 2017). If the bulk of improvements had been domain general then we would expect similarly sized improvements on other tasks measuring visuospatial WM (i.e. Mr X), but these improvements were relatively small. This is further emphasised when we look at the size of improvements relative to those in the control group, indicative of practice effects. These changing task relationships underscore the fact that the cognitive processes we recruit for individual tasks are not necessarily static but are instead subject to change as a function of experience.

Whilst most of the correlational relationships pertaining to the SOM node weights remained stable across training, those between Mr X-Backward Digit and between Forward Digit-Backward Digit, changed significantly. The correlation between the Mr X-Backward Digit pair decreased substantially following training, whereas there was a moderate increase in the correlation between the Forward Digit-Backward Digit pair. Again, this shows that relationships between tasks, as represented by the SOM, are subject to change following training. One possibility is that as subjects practice the Backwards Digit task – a version of which exists in the training battery – they gradually start to recruit similar cognitive processes or strategies that they previously used for the Forward Digit task, like chunking. The end result is that the SOM represents these tasks more similarly following training. By contrast, the Backward Digit task is now represented less similarly to the other complex span task in the assessment battery, Mr X. In short, even though both Backwards Digit and Mr X are described as WM tasks, and both improve overall following training, the change in the way that they are represented by the SOM indicates that different cognitive processes or strategies are recruited for them following training.

2.6.3 Subgroups with different training profiles

There is increasing interest in individual differences in cognitive training effects. The approach typically taken is to explore the impact of known factors, such as age (Schmiedek et al. 2010; Borella et al. 2014), baseline ability (Guye et al., 2017; Bürki et al. 2014; Zinke et al. 2014) or cognition-related beliefs (e.g. malleability of intelligence; Jaeggi, 2014), on training related gains. Here individual differences were examined in a different way, by identifying subgroups in the training profiles themselves. Clustering identified four groups that differed in their performance across tasks (High, Medium (visuospatial and verbal profiles), and Low). Changes in group membership following training were associated with the magnitude, and patterns of, gain scores. This suggests there are differential improvement trajectories among children, which would be lost in conventional group-level comparisons. These improvement profiles were meaningfully associated with fluid intelligence: those who made the largest improvements across all measures (movers to the highest performing group) had significantly higher fluid reasoning skills compared with those who stayed in one of the low-performing groups. General intelligence is thought as the ability to reason and solve novel problems (Duncan & Owen, 2000), or an index for flexible cognitive resources believed to play a critical role in the process of decomposing the unfamiliar tasks into their component parts (Duncan et al., 2017). This may indicate that the ability to abstract and

generalise newly-learned routines to unpractised tasks is one of the deciding factors of transfer effects (Gathercole et al, 2019).

The positive association between fluid intelligence and improvement profile is reminiscent of some previous studies that have shown age-related and ability-related magnification effects in the context of cognitive training (e.g., Bürki et al. 2014; Guye et al., 2017). Magnification effects are more typically observed in the context of strategy-based training than process-based training (e.g., Karbach and Verhaeghen 2014; Karbach et al., 2017), possibly indicating that the training intervention in this study facilitated strategy acquisition (Guye and von Bastian, 2017). Indeed, it has been shown that training-related improvements in working memory may be mediated by implicit development of task-specific strategies such as grouping of sequential information for recall (Dunning & Holmes, 2014; Minear et al., 2016). Gathercole et al (2019) argue that these kinds of effects are evidence that training-related gains rely on the construction and refinement of new cognitive routines and strategies. Individuals with higher levels of cognitive performance at baseline may have more capacity to acquire and perform strategies that enhance the training effect (Lövdén et al. 2012). Our findings would support this. An interesting line of enquiry would be to investigate whether children with relatively low intelligence scores could benefit from explicit instructions to help aid strategy generation while training.

2.6.4 Summary

Several task relationships changed following training (according to two separate measures), thereby indicating that the underlying mechanisms tapped by training might be task-specific rather than domain-general, and subject to change over time. Moreover, task performance trajectories were subject to individual differences under this paradigm. These findings highlight the need to reconsider the interpretation of training-related gains. Children could improve significantly on a particular task via learning specific strategies whilst having moderate or no gains on other tasks claimed to measure the same construct (Moreau et al., 2016; Gathercole et al., 2019).

To remedy this, previous studies investigating the training-induced improvement on the ability level used latent factor analysis, which is necessarily constrained by how the observed variables load onto the latent factors before and after the training for the sake of model comparability and interpretability (Schmiedek et al., 2010; Lövdén et al. 2012; Karbach et al., 2017; Guye et al., 2017; Bürki et al. 2014). However, this assumption is

challenged by the current findings, which imply that training does not only enhance performance, but also alters task structures. In Appendix A, by fitting linear models to the data, I show that this is indeed the case in the context of the current dataset. The difference in best fitting model before and after training could either be due to the enhancement of task-specific processes, an increase in individual variance across tasks, or both. Either way, it suggests that the best latent variable model before and after the training may not necessarily be the same. Fitting different models pre- and post-training would limit the meaningfulness of comparisons across time points (Dimitrov, 2006). Conversely, imposing parameter invariance when the real data suggests otherwise could lead to a large estimation bias of the model, which cannot be reliably indicated by fit statistics (Clark et al., 2018). If such cases arise, the SOM approach taken here is a potentially more flexible alternative that does not rely on as many assumptions, while still allowing for meaningful comparisons over time.

2.6.5 Limitations

Importantly, the findings here may be specific to the set of training and assessment tasks that were available. Moreover, the dataset was a composite of many individual studies, with independent recruitment criteria and assessment protocols, potentially introducing additional variability. Some have argued that training across a broad range of tasks may bring about more generalisable and enduring benefits (Green & Bavelier, 2008; Klingberg, 2010). However, broad training regimes make it difficult to identify the precise mechanisms responsible for training/transfer effects. This is because any transfer effects may be due to multiple different processes affected or produced by the training (Holmes et al., 2019; Norris et al., 2019; Smid et al., 2020). Moreover, the assessment tasks used here were not explicitly and systematically varied relative to the training tasks, further compromising the interpretability of the training/transfer effects. Nonetheless, the primary aim was to demonstrate a proof of principle, with potential benefits for those exploring multivariate profiles of change. The next step is for this to be tested in well-powered training studies with different types of assessment tasks and ranges.

2.6.6 Conclusion

SOM models provide an effective alternative for the representation and prediction of multivariate data typically found in training studies. Applying SOMs to the current data revealed nuanced task relationships that are subject to change following working memory training, suggesting that the underlying mechanisms of improvement may be task-specific

rather than domain-general. The use of K-means clustering revealed distinct subgroups with differentiable improvement trajectories. These improvement trajectories were related to pre-training fluid intelligence.

Chapter 3: Training and transfer within nested tasks: a perceptual discrimination paradigm

3.1 Introduction

Training on one or more cognitive tasks can improve performance on other, un-practised, tasks. In the previous chapter I used a novel multivariate approach to explore transfer and individual differences following a broad training regime across a set of memory tasks. The training resulted in substantial transfer to the assessment tasks, relative to controls. K-means clustering identified subgroups in terms of performance profiles and changes in group membership following training were at least partially mediated by pre-training fluid reasoning ability. The training also resulted in significant changes to task relationships both with respect to the between task correlations and multivariate relationships as represented by the SOM model. In the previous study I investigated a broad training programme. Whilst in theory this may encourage broader transfer effects, in practice this does not appear to be the case (Green & Bavelier, 2008; Klingberg, 2010; Sala & Gobet, 2019; Simons et al., 2016). Moreover, as touched upon in the previous chapter, training on a range of tasks makes it hard to pin down precisely the mechanisms responsible for training gains and/or transfer, or a lack thereof. This is because any transfer effects may stem from a multiplicity of processes affected by, or brought about by, the training. This is especially true when the assessment tasks are not systematically related to the training tasks (Holmes et al., 2019; Norris et al., 2019).

In this chapter, I wanted to move away from using existing datasets and broad training regimes, and instead implement an experimental approach in which the training regime was specific, and the training tasks were systematically related to the assessment tasks in a fine-grained manner with respect to task features. Tasks necessarily vary along several dimensions (e.g. stimulus type, spatial properties, timings, and goals) and there are multiple ways to calculate and conceptualise task similarity (e.g. correlational, hypothetical modelling, task analytical), each of which possess their own strengths and weaknesses. Correlational approaches are a convenient data driven way of defining task similarity but are unstable and by themselves cannot be used to make causal inferences (Gogtay & Thatte, 2017; Smid et al., 2020; Kievit et al., 2011; Maul et al., 2016). Modelling approaches to establishing task similarity such as production or connectionist style models are more stable and theoretically driven but are not readily available and rarely agreed upon (Taatgen, 2013; Yang et al,. 2019; Gathercole et al., 2019; Smid et al., 2020). Task analytical approaches require the researcher

to identify and specify the extrinsic features of a task (e.g. stimulus type, spatial properties, timings, and goals), they are relatively simple and stable but can suffer from a lack of granularity and still require mapping onto theory. Importantly, these different approaches to task similarity are not mutually exclusive and ought to be used in tandem where possible. Nonetheless, to explain why transfer tends to be limited in scope and to establish more concrete boundary conditions, the field is now calling for more tightly controlled, higher resolution, and systematic manipulations of extrinsic task features in high powered studies (Katz et al., 2017; Redick, 2019; Sala & Gobet, 2019; Von Bastian & Oberauer, 2014; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Taatgen, 2013). This is particularly important given that transfer tends to be tied to specific task features (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019).

Towards this aim, this chapter examines how feature overlap informs transfer within a set of hierarchically nested visual-discrimination tasks. The idea being that this would allow me to better identify specific mechanisms and potential boundary conditions for transfer by exploring the impact of various types of feature relationships between the training and assessment tasks. Moreover, by organising the tasks hierarchically I was able to ask whether the direction of transfer cascades in a unidirectional or bidirectional manner relative to the position of the training task within the hierarchy.

3.1.1 Transfer specificity in discrimination and switching tasks

Most of the cognitive training research to date has focused on relatively higher order tasks with the hope that these would have more generalisable benefits (Melby-Lervag et al., 2016; Simons et al., 2016), although their specificity is becoming increasingly apparent. In contrast, the specificity of lower order tasks is already well established (Fahle, 2005). Transfer in simple visual discrimination tasks is often confined to the specific stimulus features (e.g. orientation, contrast, motion) and contexts (range, spatial location, category etc.) being trained (Dosher & Lu, 2017; Fahle, 2005). However, this is still a graded phenomenon, both on-task learning and transfer to albeit very subtly different tasks, have been shown to be dependent upon stimulus complexity, judgement precision, and specifics of the test/training procedures (Assihar & Hochstien, 2004; Berry et al., 2010; Dosher & Lu, 2009; Dosher & Lu, 2017; Fahle, 2005; Jacobs, 2002; Jeter et al., 2009; Jeter et al., 2010; Parsons et al., 2016). Aside from the well demonstrated signal to noise ratio improvements of representations in visual processing brain regions following training, enhanced perceptual discrimination ability is also thought to stem from modifications in top-down attentional and

decision-making processes, providing another potential avenue for transfer to manifest (Assihar & Hochstien, 2004; Berry et al., 2010; Covey et al., 2018; Dosher & Lu, 2017; Lu & Dosher, 2009; Parsons et al., 2016).

A paradigm in which the specificity of training induced transfer is less clear is that of simple task-switching. In line with other paradigms, task-switch training, wherein participants practise switching between two (or more) simple binary decision tasks, shows fairly consistent transfer to other similarly structured switching tasks involving different binary decisions (Dorrenbacher et al., 2014; Karbach & Kray, 2009; Minear et al., 2002; Minear & Shah, 2008; Zinke et al., 2012). However, Karbach & Kray (2009) found that task-switch training improvements also transferred to interference control tasks (color-stroop/number-stroop), verbal working memory tasks (reading-span/counting-span), spatial WM tasks (symmetry-span/navigation-span), and most surprisingly fluid intelligence tasks (figural reasoning/letter series/ravens matrices). They suggest that improved executive control processes tapped by the switching paradigm, such as interference control, are common across these tasks and may thus be responsible for these findings. Similar studies (Dorrenbacher et al., 2014; Zinke et al., 2012) have failed to show any generalizable benefits to interference control/inhibition tasks (flanker, stroop), nor updating/working memory tasks (n-back, keep track, backward-digit-span, counting span) but did find some evidence for transfer to a measure of processing speed (choice reaction time). Discrepancies between studies could be due (amongst others) to differences in sample populations, training/assessment task specifics, training dosages, motivation, sample sizes, and analysis protocols (Dorrenbacher et al., 2014; Karbach et al., 2017; Simons et al., 2016; Zinke et al., 2012). Given these mixed findings it is difficult to ascertain the scope of transfer for task-switching training.

The concept of a task-set mentioned in Chapter 1 may shed light on the specificity of transfer in task-switching contexts, and perhaps more generally. One popular account is that task-sets perform a shielding function by providing a preparatory attentional state that serves to bias the set of imminent processes recruited to perform a task, so as to prevent irrelevant stimulus features, or indeed stimulus-response mappings, from interfering with the correct response process (Rogers & Monsell, 1995; Dreisbach & Wenke, 2011). However, the adoption of specific task-sets in a switching context may become proximately maladaptive; switch-costs arise because the need to switch between task-sets slows people down and/or a failure to adequately switch brings about pro-active interference caused by an irrelevant task-

set (Dreisbach & Wenke, 2011). Interestingly, task-switching training appears to relax task-set shielding. This is evidenced by the finding that irrelevant information for both tasks can interfere with performance in a task-switching context but not on their single task counterparts (Dreisbach & Wenke, 2011). This implies that prior exposure to a task through training (or simply by task-order) may initially cause some negative transfer effects on a different task. Moreover, the broader transfer observed to other tasks involving interference control (Karbach & Kray, 2009) may be due to the adoption of more relaxed task-sets relative to training on single task-counterparts. Further support for this comes from Sabah et al (2019) who found that increasing task-variability (in terms of content and structure) in a task-switching context resulted in greater transfer to novel switching tasks.

3.1.2 Motivation for current study

Recent theoretical and experimental work suggests that specificity is the rule rather than the exception for transfer effects (Melby-Lervag et al., 2016; Sala & Gobet, 2019; Simons et al., 2016; Gathercole et al., 2019). However, the specificity of transfer varies between task-paradigms and as a function of task complexity and novelty (Assihar & Hochstien, 2004; Dosher & Lu, 2017; Gathercole et al., 2019; Jeter et al., 2009; Taatgen, 2013). To compliment theoretical progress and better understand the precise nature of transfer and its boundary conditions, further experimental studies are required that identify and systematically manipulate the overlap in task features between training and assessment tasks (Gathercole et al., 2019; Holmes et al., 2019; Minear et al., 2016; Norris et al., 2019; Von Bastian & Oberauer, 2014). Visual discrimination and task-switching paradigms both show potential for transfer (Assihar & Hochstien, 2004; Dorrenbacher et al., 2014; Dosher & Lu, 2017; Fahle, 2005; Karbach & Kray, 2009). Moreover, their simple feature structures (see figure 3.2), for example with switching elements being added to impose executive demands, make them suitable for the systematic exploration of practice induced transfer effects.

3.1.3 The present study

The present study explored the potential transfer of training two tasks within a set of six hierarchically nested perceptual discrimination tasks. To do so, I conducted a large online training study powering for small-medium effect sizes. The tasks were hierarchically nested with respect to their combination of task features. That is, the higher-level tasks contain *all the features* of their lower-level counterparts. Importantly, the focus here was on task features and I do not make strong claims about the specifics of the associated cognitive processes – it

is very difficult to infer a cognitive hierarchy, especially for bespoke tasks. However, I assume that these tasks span a range of processes, including: attention, working memory, and executive control (Dosher & Lu, 2017; Miyake et al., 2000), and that as tasks contain more features, so the required cognitive processes become more complex. Whilst the stimuli were identical, the task features varied systematically with respect to judgement type (number of spikes or 'spikiness'), presentation type (simultaneous or delayed) and task-switching, allowing them to be established as potential boundary conditions. All participants completed each of the six assessment tasks both before and after training. Participants were randomised to three training groups: one group received training on a relatively low-level task, another group received training on a relatively high-level task, and a third group trained on a control task.

There were several motivations for taking this approach: 1) Task overlap can be quantified in an unambiguous and systematic manner at the level of the task features; 2) Given the prevalence of transfer specificity, I wanted the variability between tasks to be fairly minimal, systematic and precise, to allow for transfer; 3) Relatedly, transfer seems to depend upon complexity/novelty, so I chose tasks that were relatively simple to interpret and easy to learn, whilst still being complex and novel enough to allow for transfer; 4) The simplicity of the tasks and the brevity of their trials made for a relatively parsimonious and cost-effective study, allowing me to collect a large sample.

Despite potential avenues for transfer, I was hesitant to make specific a-priori predictions given the novelty of the specific task parameters, stimuli, and training protocols used. Instead, I posed the following open questions: 1) Do participants make substantial on-task training gains? 2) Do the different training tasks generate different transfer patterns? 3) Are these transfer patterns predicted by the proportion of overlapping features? 4) Do some shared features contribute more to the transfer than others? 5) Is the direction of transfer unidirectional or bidirectional relative to the hierarchical position of the training task?

3.2 Materials and methods

3.2.1 Ethical approval

This study received ethical approval from the Cambridge Psychology ethics committee, University of Cambridge, application number: PRE.2019.046. All participants provided informed consent by checking a box to confirm they had fully understood the implications of participation and their right to withdraw.

3.2.2 Participants

The final sample (see 'Data Exclusion') consisted of 175 English speaking adults with normal/corrected vision aged between 18 and 35 years of age (*M*=27.11, *SD*=4.85). Participants were recruited via 'Prolific', a platform for recruiting and paying people to participate in online experiments. Participants were paid at a rate of £6 per hour and received a £5 bonus upon completion of all sessions.

A total sample size of 175 in three groups yielded 0.84 power to detect a medium transfer effect size (d = 0.5). Participants were randomly assigned to three groups and their demographics are displayed in Table 3.1. Analyses revealed moderate evidence for no group differences with respect to age, (F(2,172)=1.23, p=0.293, BF_{10}=0.16, and gender, X^2=2.31, df=2, p=0.314, BF_{10}=0.14.

Table 3.1. Group demographics.

	SSPT	DSWT	Control
N	59	60	56
Age: *M* (*SD*)	26.45 (5.28)s	27.05 (4.84)	27.87 (4.34)
Female: *N* (%)	30 (50.8%)	29 (48.3%)	21 (37.5%)
Male: *N* (%)	29 (49.2%)	31 (51.7%)	35 (62.5%)

Note. Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT).

3.2.3 Stimuli

A set of 220 (20x11) spikey shapes was generated using MATLAB as specified by Van Dam & Ernst (2015). The shapes varied in a graded fashion along two dimensions: 'spikiness' and 'number of spikes' (see Figure 3.1). They were always the same turquoise-grey on a black background. The range of both the Number of Spikes and Spikiness dimensions was determined from task pilot data. Seven difficulty levels were chosen to capture the range of performance and to allow room for improvement. Task difficulty corresponded to the deviation between stimuli along either dimension, where a difference of one was the most difficult judgment and a difference of seven was the easiest judgment.

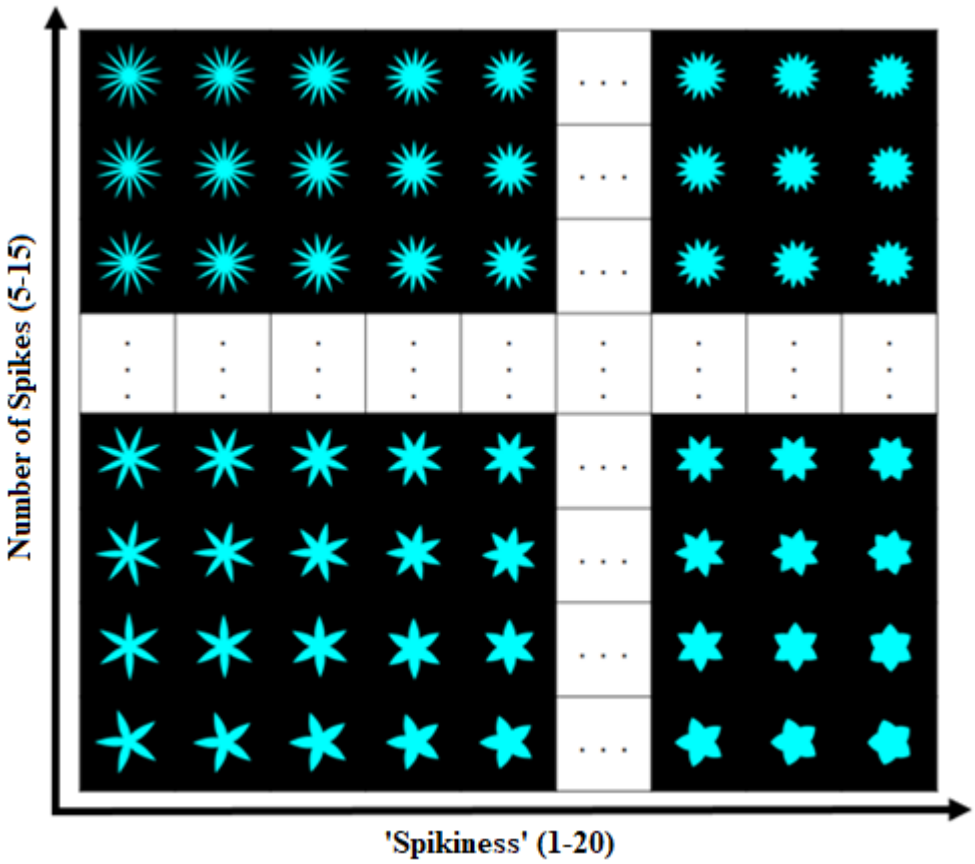
All tasks were coded using JavaScript (jsPsych; De Leeuw, 2015), HTML, and CSS in house. I used JATOS to set up and run the study on a local server.

3.2.4 Assessment tasks

In each assessment phase (pre- and post-training) participants completed two blocks on each of the six tasks. Task order was semi-randomised with the constraint that two non-

switching tasks had to occur first. All six assessment tasks required the participant to make a same-different judgement about two spikey shapes (see Figures 3.2 & 3.3). Participants were instructed to press the ‘J’ key when making a same-response or the ‘F’ key when making a different-response. They were instructed to simply be as accurate as possible, and no mention of speed was made. This was to avoid ambiguously introducing a large range of potentially viable speed accuracy trade-offs and aimed at making the results more interpretable.

Figure 3.1. The stimulus set comprised 220 spikey shapes that varied in a graded fashion along two dimensions: ‘spikiness’ and ‘number of spikes’.



Each block contained 56 trials, half required a ‘same’ response, the other half required a ‘different’ response, and these were evenly distributed across the seven difficulty levels (eight trials at each difficulty). All participants saw the same stimuli as one another within every task but in a randomised order. Participants received explicit step by step instructions with examples for each task, along with a small number of practice trials. Feedback was provided on each trial, with the shape turning green for 300ms to indicate ‘correct’ and red for ‘incorrect’. There was a 200ms inter-trial interval. Participants received feedback about their average accuracy after each block. Each assessment phase took approximately 45mins.

The tasks were divided into those with the two shapes presented simultaneously and those with the two shapes presented sequentially. In tasks using simultaneous presentation,

participants were shown a centred fixation cross for 350ms followed by two spikey shapes presented simultaneously alongside one another for 1600ms, they had to respond within the 1600ms else the trial was counted as incorrect. For tasks using sequential presentation, participants were shown a fixation cross for 350ms, followed by a 'target-spikey-shape' for 800ms, which was then masked for 1000ms, and followed by a second 'response-spikey-shape' for 1400ms. Participants had to respond within the 1400ms otherwise the trial was counted as incorrect. The simultaneous presentation tasks were set to have slightly longer response deadlines to account for the increased encoding demands during the response phase of the task (two stimuli vs one).

The tasks were further sub-divided into those that required participants to make judgements about the 'Spikiness' property of the shape, those that required participants to make judgements about the 'Number of Spikes' property of the shape, and those that required participants to switch between these two judgement types. When a Spikiness-judgement was required, the two shapes always shared the same number of spikes (varying randomly between 5 and 15 spikes) and participants were to make a judgment about whether the two shapes share the same 'Spikiness' or not. When a Number-of-Spikes-Judgement was required the two shapes varied in both their 'Spikiness' and their number of spikes and participants were to make a judgment about whether the two shapes share the same number of spikes or not. When switching between judgements, the colour of the border in the response phase cued the judgement dimension ('Spikiness' or 'Number of Spikes'). A blue border cued the 'Spikiness' judgement and a red border cued the 'Number of Spikes' judgement. This provided the following six assessment tasks (see below for individual task descriptions and Figure 3.2 for a graphical depiction): Simultaneous-Spikiness (SSP); Simultaneous-Number (SN); Simultaneous-Switching (SSW); Delayed-Spikiness (DSP); Delayed-Number (DN); Delayed-Switching (DSW). Crucially, these tasks are all hierarchically nested, with the more complex variants being formed of their constituent paradigms.

Simultaneous-Spikiness (SSP)

Participants are shown a fixation cross for 350ms followed by two spikey shapes presented simultaneously alongside one another for 1600ms. In this task, the two spikey shapes always share the same number of spikes and participants are required to make a judgment about whether the two shapes share the same 'spikiness' or not.

Simultaneous-Number (SN)

Participants are shown a fixation cross for 350ms followed by two spikey shapes presented simultaneously alongside one another for 1600ms. In this task, the two spikey shapes can vary in both their 'spikiness' and their number of spikes and participants are required to make a judgment about whether the two shapes share the same number of spikes or not.

Simultaneous-Switching (SSW)

Participants are shown a fixation cross for 350ms followed by two spikey shapes presented simultaneously alongside one another within a border for 1600ms. In this task, the colour of the border cues the participant as to which judgement dimension ('spikiness' or number of spikes) they ought to be responding along on a given trial. If the border is Blue, the two spikey shapes always share the same number of spikes and participants are required to make a judgment about whether the two shapes share the same 'spikiness' or not. If the border is Red, the two spikey shapes can vary in both their 'spikiness' and their number of spikes and participants are required to make a judgment about whether the two shapes share the same number of spikes or not.

Delayed-Spikiness (DSP)

Participants are shown a fixation cross for 350ms followed by a target-spikey-shape for 800ms, then a masked delay of 1000ms, then a second response-spikey-shape for 1400ms. In this task, the two spikey shapes always share the same number of spikes and participants are required to make a judgment about whether the target and response stimuli share the same 'spikiness' or not.

Delayed-Number (DN)

Participants are shown a fixation cross for 350ms followed by a target-spikey-shape for 800ms, then a masked delay of 1000ms, then a second response-spikey-shape for 1400ms. In this task, the two spikey shapes can vary in both their 'spikiness' and their number of spikes and participants are required to make a judgment about whether the target and response stimuli share the same number of spikes or not.

Delayed-Switching (DSW)

Participants are shown a fixation cross for 350ms followed by a target-spikey-shape for 800ms, then a masked delay of 1000ms, then a second response-spikey-shape within a

border for 1400ms. In this task, the colour of the border cues the participant as to which judgement dimension ('spikiness' or number of spikes) they ought to be responding along on a given trial. If the border is Blue, the two spikey shapes always share the same number of spikes and participants are required to make a judgment about whether the target and response stimuli share the same 'spikiness' or not. If the border is Red, the two spikey shapes can vary in both their 'spikiness' and their number of spikes and participants are required to make a judgment about whether the target and response stimuli share the same number of spikes or not.

3.2.5 Training tasks

Training was conducted on either the Simultaneous-Spikiness task (SSPT), Delayed-Switching task (DWST) or a Speeded-Response-Mapping task (Control; see Figure 3.2). These represented tasks relatively low in the hierarchy, relatively high in the hierarchy and a control, respectively. Participants received three sessions of adaptive training with eight blocks per training session and 20 trials per block. As in the assessments, all training tasks had seven difficulty levels. Participants started at the easiest difficulty level on the first session, difficulty was then adapted at the end of each block, and the level reached by the end of each session carried over into the next training session. Level up/down performance requirements were based on preliminary pilot data and aimed at generating somewhat similar improvement trajectories over time across the training groups.

Simultaneous-Spikiness-Training (SSPT)

This training task is identical in structure to the Simultaneous-Spikiness assessment task. Each training session lasted approximately 15mins. If participants achieved >75% accuracy they moved up a difficulty level, if they achieved <65% accuracy they moved down a difficulty level, otherwise they remained at the same difficulty level.

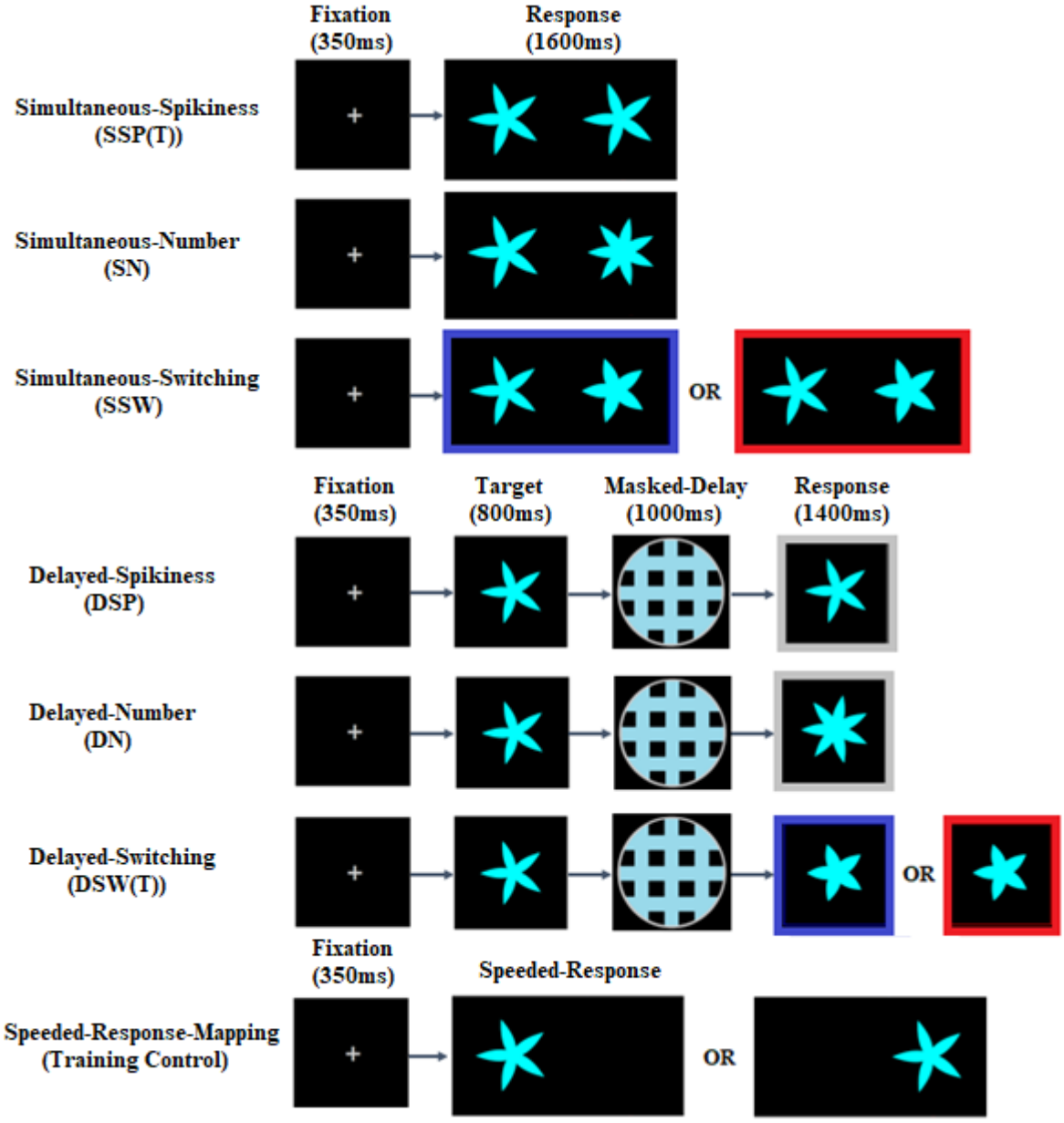
Delayed-Switching-Training (DSWT)

This training task is identical in structure to the Delay-Switching assessment task. Each training session lasted approximately 20mins. If participants achieved >65% accuracy they moved up a difficulty level, if they achieved <55% accuracy they moved down a difficulty level, otherwise they remained at the same difficulty level.

Speeded-Response-Mapping-Training

In the Speeded-Response-Mapping-Training task participants are shown a fixation cross for 350ms followed by a spikey shape in one of two locations (left or right) and are required to press the key corresponding to the location of the stimulus ('F' for left and 'J' for right) as quickly as they can. The difficulty was adjusted by changing the limited amount of time participants had to make a response. There were seven difficulty levels: 550ms, 500ms, 450ms, 400ms, 350ms, 300ms, 250ms, and 200ms (these times were chosen based on data from the human benchmark project (https://www.humanbenchmark.com/tests/reactiontime). Participants receive feedback about whether or not they made the correct choice: Green for correct and Red for incorrect (300ms). A failure to respond within the time limit is counted as incorrect. There was a 200ms inter-trial interval. Each training session lasted approximately 12mins. If participants achieved >90% accuracy and their reaction time was less than the current difficulty level they moved up a difficulty level, otherwise they moved down a difficulty level.

Figure 3.2. Training and assessment task trial sequences.

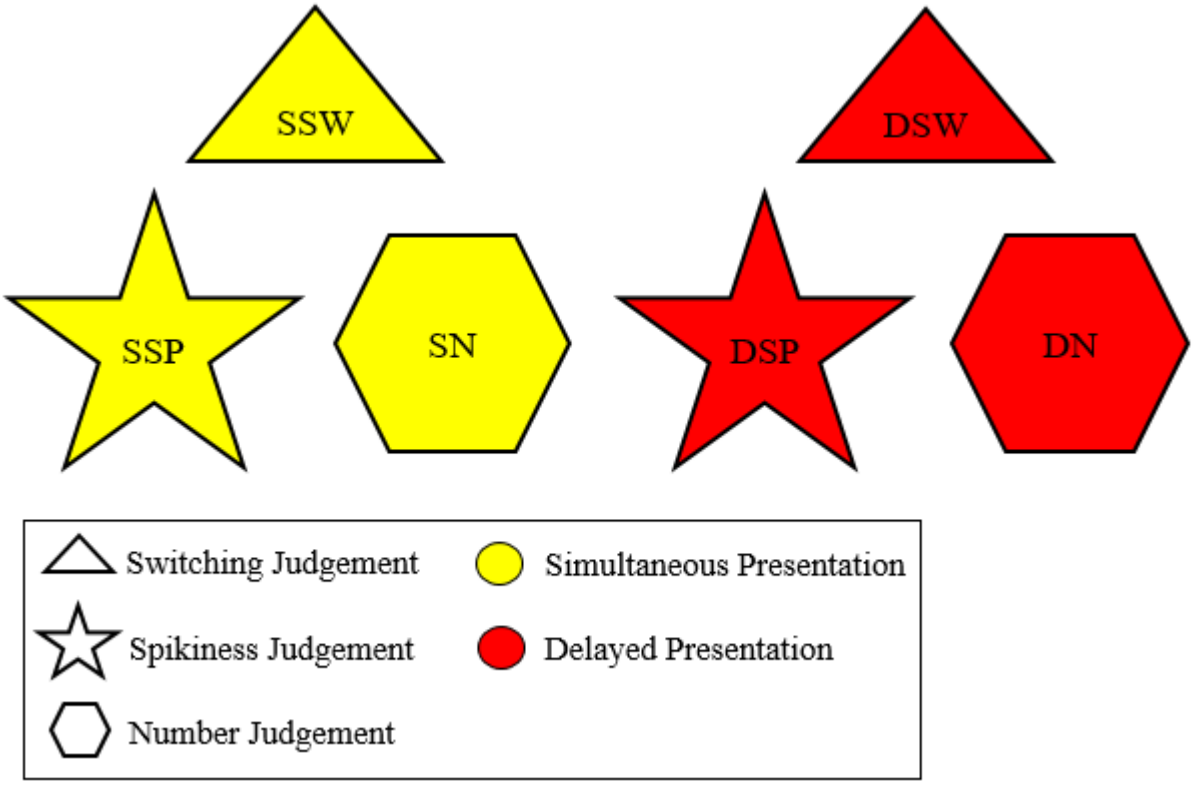


Note. All tasks were used during assessment except for the Speeded-Response-Mapping task, which was only used in training for the Control group. The additional (T) indicates that the task was both an assessment task and training task.

3.2.6 Training procedure

After the Pre-Training assessment, participants were randomly allocated to one of the three training conditions and received specific instructions about the training phase along with a personalised 'homepage'. This homepage included the number of training sessions completed and how long they had to wait before starting the next session. Participants were only allowed to start the next session after 10 hours had elapsed from the previous. On the training homepage there was also a link to the post-training assessment session that they could access 10 hours after completing all the training sessions.

Figure 3.3. Depiction of the task feature hierarchy

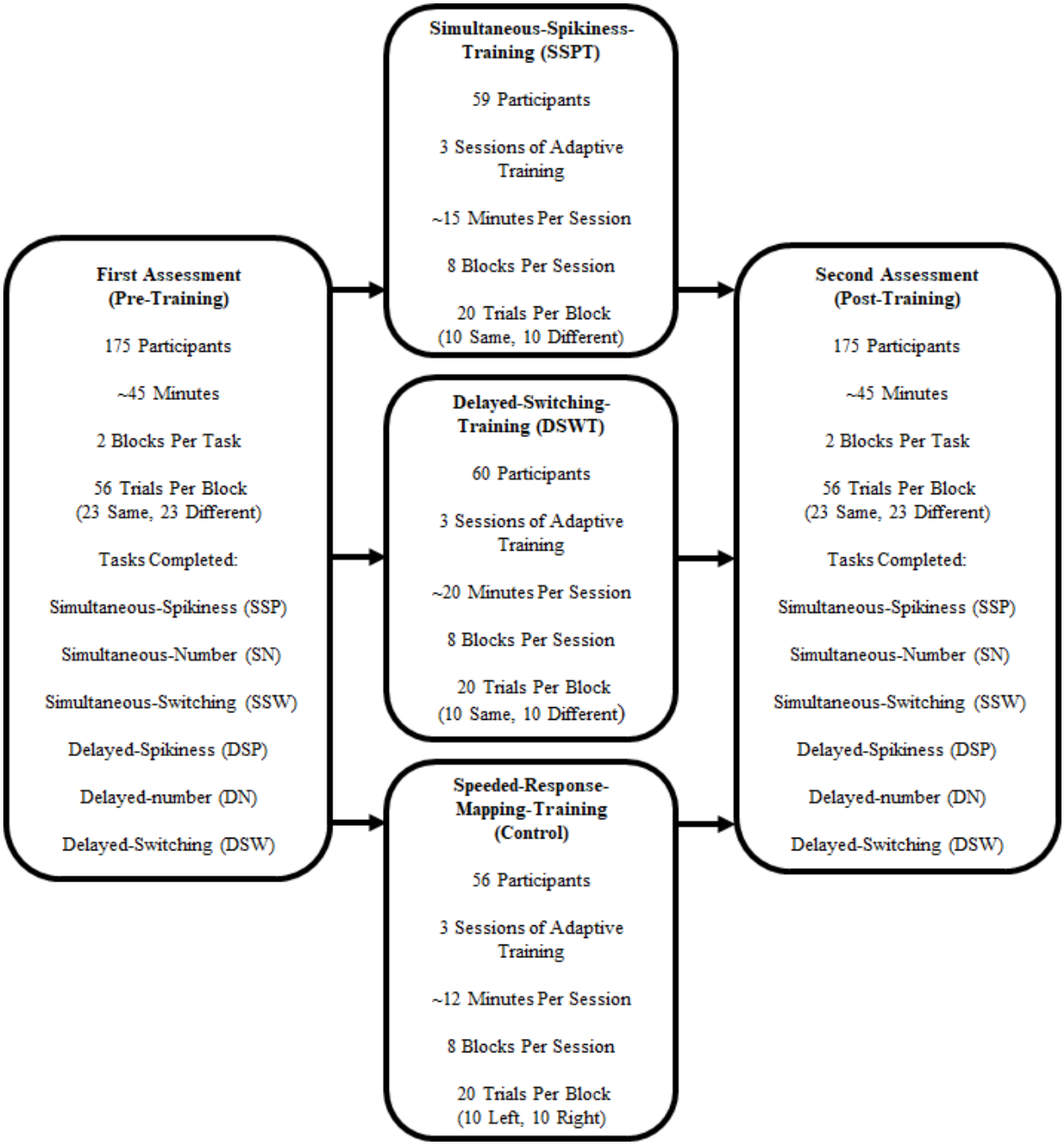


Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW).

3.2.7 Overview of procedure

All participants signed up and completed all sessions online via Prolific. All participants completed the same set of six assessment tasks, each of which required the participant to make same-different judgements about two spikey shapes, both before (pre) and after (post) training: Simultaneous-Spikiness (SSP); Simultaneous-Number (SN); Simultaneous-Switching (SSW); Delayed-Spikiness (DSP); Delayed-Number (DN); Delayed-Switching (DSW). Upon completion of the first assessment session participants were then randomly allocated to one of three training groups: Simultaneous-Spikiness-Training (SSPT); Delayed-Switching-Training (DSWT); or Speeded-Response-Mapping-Training (Control). The first two groups (SSPT and DSWT) trained on their assessment task counterparts (SSP and DSW), whilst the third group trained on alternative task that acted as a control. Each training group then received three sessions of adaptive training spaced out across a few days before completing the second assessment session (see figure 3.4).

Figure 3.4. Overview of Procedure.



3.2.8 Data exclusion

All incoming data were screened for quality based on summary statistics saved using JavaScript/JATOS. Participants with particularly low accuracy and reaction times across tasks at pre-training assessment (Accuracy<56% and RT<600ms; based on pilot data) were assumed to not be engaging and excluded from the study. Furthermore, participants who did not complete all sessions were excluded from analysis. Of the 199 participants who started, 183 participants completed all sessions.

After data collection, participants who scored below 2 standard deviations (calculated task wise at pre-training) on two or more tasks at pre- or post-training were excluded from all subsequent analyses. Again, this was intended to remove participants who were not engaging with the tasks. This resulted in 8 out of the 183 participants to be excluded (Simultaneous Spikiness Training=5, Delayed Switching Training=1, Control=2; Chi-Square:X^2=3.25, p= 0.196, BF_{10}= 0.867), leaving 175 in total (59 in the Simultaneous Spikiness Training group, 60 in the Delayed Switching Training, and 56 in the Control group).

Further to this, univariate data points 1.5 times the interquartile range above the third quartile or below the first were considered statistical outliers and individuals were excluded from any analyses on the respective task. This resulted in 6, 5, 1, 3, 7, and 4 of the 175 participants to be excluded from tasks Simultaneous Spikiness, Simultaneous Number, Simultaneous Switching, Delayed Spikiness, Delayed Number, and Delayed Switching respectively.

Training data were partially missing for 17 of the participants (Simultaneous Spikiness Training=4, Delayed Switching Training=8, Control=5). To my knowledge they completed the training session, and the missing data was caused by an unknown technical issue when attempting to upload their data to the server. As such, these participants were removed from the training data analyses but still included in the rest of the analyses.

3.3 Analysis plan

Data were analysed using both traditional null-hypothesis significance testing (NHST) and the more recently advocated Bayesian methods. Statistics are reported for both methods where they have been applied; however, Bayesian metrics are preferred as they allow the strength of evidence in favour of the null and alternative hypotheses to be quantified in an unbiased manner (Wagenmakers et al., 2018). All main analyses were conducted using JASP software (JASP Team, 2019). Inverse Bayes factors (BF_{10}) expressing the odds of the alternative hypothesis relative to the null are used throughout (Jeffreys, 1961; van Doorn et al., 2019). For the NHST analyses, Holm-corrected p-values with a family wise alpha of 0.05 are used throughout to adjust for multiple comparisons. Holm-correction is a slightly less conservative alternative to the Bonferroni method (see Chen et al., 2017 for more details). First p-values are ranked from smallest to largest, starting with the smallest and continuing in a stepwise fashion, the original values are adjusted according to the total number of comparisons (the greater the number of comparisons the greater the adjustment) and their

rank (the lower the rank the smaller the adjustment). Let p' = adjusted p-value, p = unadjusted p-value, α' = adjusted alpha, α = family wise alpha, m = number of comparisons, and i = 1,…,m, then:

$$p_i' = p\left(\frac{1}{\alpha_i'/\alpha}\right) \; where \; \alpha_i' = \frac{\alpha}{m - i + 1}$$

The procedure stops as soon as the first $p_i' > \alpha$ is observed and thereafter all remaining p-values are declared non-significant.

To evaluate transfer effects, I report results from ANCOVA models for each task and each group contrast, wherein post-training performance is the dependent variable, group is the independent variable and pre-training performance is a covariate. I opted for ANCOVAs instead of repeated measures ANOVAs as they are considered more powerful and less biased in randomised studies such as this one (Senn, 2006; van Breukelen, 2006). Moreover, including pre-training performance as a covariate controls for potential aptitude by treatment effects (Karbach et al., 2017). However, it is important to note that this deviates from the pre-registration analysis plan, in which I proposed using repeated measures ANOVAs (osf.io/36ayf).

3.4 Results

Task performance was primarily operationalised as accuracy because I instructed participants to be as accurate as possible within the time constraints and made no mention of speed. However, reaction time results are presented in Appendix B (Tables B.3 & B.4). Table 3.2 provides summary statistics for pre- and post-training accuracy and their differences.

3.4.1 Pre-training performance

A series of one-way ANOVAs tested for pre-training differences in task performance. I found moderate evidence only for a difference between groups on the Simultaneous-Switching task at pre-training assessment (F(2,171)=4.77, p = 0.010, BF_{10}=3.46, η_p^2=0.05). Post-hoc analyses provided strong evidence that the Control group had lower accuracy than the Delayed Switching Training group on the Simultaneous Switching task at pre-training assessment (t(113)=3.08, d=0.56, p = 0.007, BF_{10}=11.15). There was no strong evidence for group differences on the pre-training assessment when comparing the Control and Simultaneous Spikiness Training groups (t(112)=1.69, d=0.32, p=0.184, BF_{10}=0.74) nor

when comparing the Simultaneous Spikiness Training and Delayed Switching Training groups ($t(117)=1.40$, $d=0.26$, $p=0.184$, $BF_{10}=0.48$).

Table 3.2. Assessment summary statistics for accuracy performance.

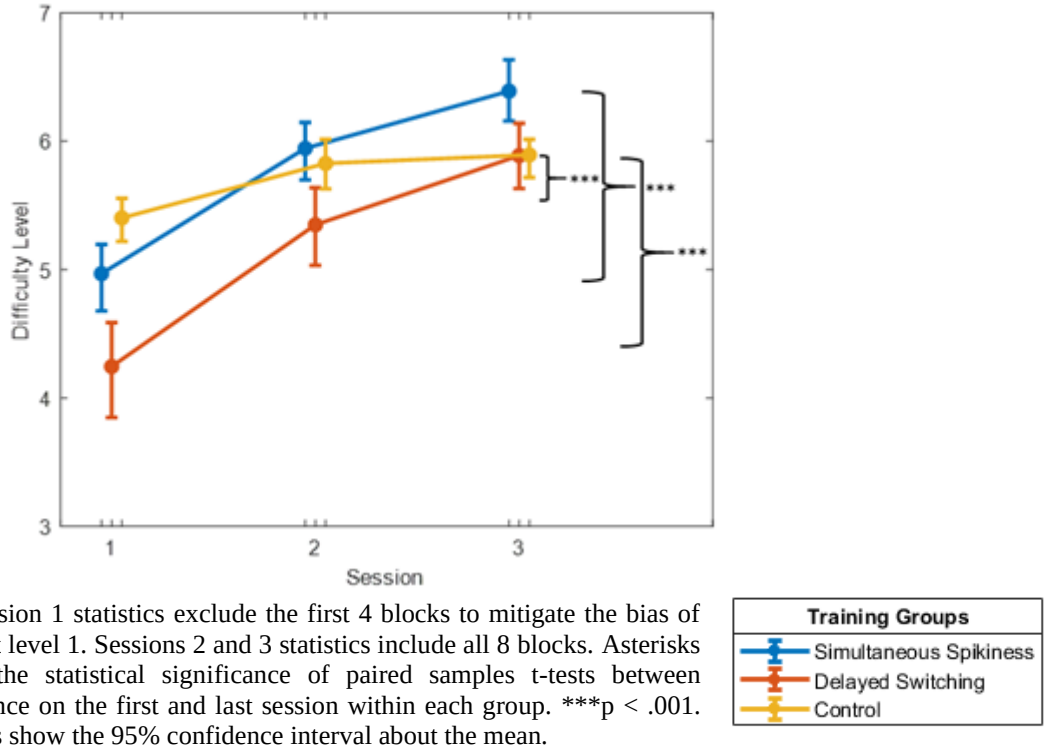
Tasks		Accuracy (%)						Paired t-test results				
Assessment	Training	Pre-training		Post-training		Difference (Post-Pre)						
		M	*SD*	*M*	*SD*	*M*	*SD*	*df*	*t*	*d*	BF_{10}	*p*
SSP	SSPT	75.69	8.01	83.94	7.33	8.25	8.18	58	7.55	1.00	>100	<0.001***
	DSWT	75.77	7.10	80.46	8.58	4.69	9.64	59	3.76	0.48	63.55	<0.001***
	Control	74.70	7.88	75.29	8.30	0.59	8.84	55	0.48	0.06	0.16	1.000
SN	SSPT	79.60	4.53	79.23	7.08	-0.37	7.04	58	0.39	0.05	0.15	0.690
	DSWT	80.07	5.12	78.12	6.71	-1.95	6.86	59	2.18	0.28	1.28	0.033*
	Control	78.18	5.45	76.80	6.62	-1.38	6.21	55	1.61	0.22	0.50	0.444
SSW	SSPT	65.83	8.20	72.64	8.34	6.81	6.75	58	7.75	1.00	>100	<0.001***
	DSWT	68.04	8.71	73.20	8.05	5.16	8.09	59	4.94	0.63	>100	<0.001***
	Control	63.12	8.69	68.12	8.50	5.00	9.05	55	4.09	0.55	>100	<0.001***
DSP	SSPT	67.28	9.27	70.07	8.38	2.79	8.78	58	2.43	0.31	2.13	0.054
	DSWT	65.90	7.38	69.73	8.72	3.83	8.09	59	3.63	0.47	43.10	<0.001***
	Control	64.27	7.01	64.77	7.91	0.50	8.01	55	0.45	0.06	0.16	1.000
DN	SSPT	73.32	7.42	75.51	6.24	2.19	7.46	58	2.23	0.29	1.41	0.058
	DSWT	72.62	7.75	75.81	6.76	3.19	8.81	59	2.70	0.36	3.93	0.018*
	Control	72.40	7.62	73.94	5.95	1.54	7.54	55	1.49	0.20	0.42	0.444
DSW	SSPT	62.00	7.07	65.65	6.49	3.65	7.39	58	3.72	0.49	55.97	<0.001***
	DSWT	60.90	6.46	68.29	7.44	7.39	6.97	59	8.07	1.05	>100	<0.001***
	Control	59.39	8.01	62.76	6.74	3.37	8.09	55	3.11	0.41	10.44	0.015*

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Task-wise holm-corrected).

3.4.2 Training task gains

Paired-samples t-tests (one tailed) were performed for each of the three training groups to establish whether participants made improvements with respect to the average difficulty level achieved between the 4th and 8th block of the first session and the final training session (see Figure 3.5). All groups made substantial training gains on their respective training tasks: Simultaneous Spikiness Training (M=1.42, SD=0.97, t(54)=12.87, d=1.73, p<0.001, BF_{10}>100); Delayed Switching Training (M=1.64, SD=1.23, t(51)=9.55, d=1.32, p<0.001, BF_{10}>100); Control (M=0.49, SD=0.47, t(50)=7.41, d=1.03, p<0.001, BF_{10}>100).

Figure 3.5. Improvements on the trained task for each group across training sessions.

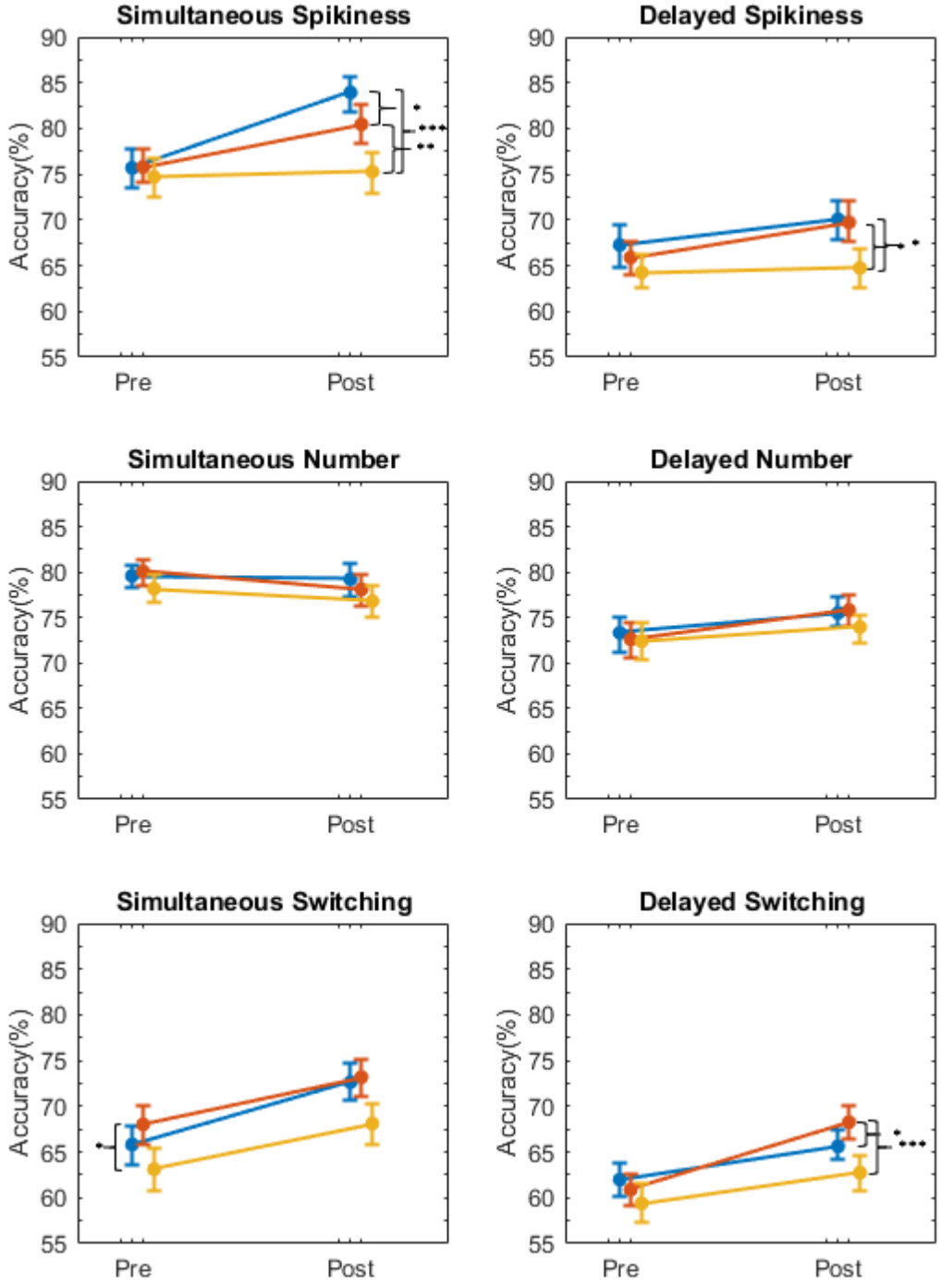


Note. Session 1 statistics exclude the first 4 blocks to mitigate the bias of starting at level 1. Sessions 2 and 3 statistics include all 8 blocks. Asterisks indicate the statistical significance of paired samples t-tests between performance on the first and last session within each group. ***p < .001. Error bars show the 95% confidence interval about the mean.

3.4.3 Transfer effects

To investigate whether the groups show differential transfer patterns, I conducted a series of ANCOVAs to establish group differences in post-training performance, whilst covarying for pre-training performance. The full results are shown alongside the corresponding descriptive statistics for each task and each group contrast in Table 3.3 and Figure 3.6. Positive evidence in support of the alternative hypothesis for group differences is summarised below. A statistical comparison of these effect sizes is also provided in the Appendix B.

Figure 3.6. Mean accuracies pre- and post-training for each group on each task.



Note. Significant group differences are shown at pre-training (from Table 3.2) and at post training after controlling for pre-training performance (from Table 3.3). Error bars show the 95% confidence interval about the mean. *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected).

Table 3.3. Pairwise group ANCOVAs of post-training accuracy adjusted for baseline performance.

Group Contrast	Task	Post-training accuracy difference (%)	ANCOVA				
			df	*F*	*p*	BF_{10}	η_p^2
SSPT-Control	SSP	8.25	(1,106)	36.264	<0.001***	>100	0.254
	SN	1.64	(1,108)	1.837	0.534	0.480	0.016
	SSW	3.01	(1,111)	5.122	0.075	2.027	0.044
	DSP	3.89	(1,110)	7.940	0.012*	7.309	0.067
	DN	1.26	(1,109)	1.418	0.472	0.383	0.012
	DSW	1.96	(1,110)	2.865	0.093	0.768	0.025
DSWT-Control	SSP	4.77	(1,110)	9.928	0.004**	16.181	0.082
	SN	0.32	(1,109)	0.076	0.782	0.213	0.000
	SSW	2.78	(1,112)	3.941	0.098	1.317	0.034
	DSP	4.07	(1,110)	8.459	0.012*	8.835	0.071
	DN	1.81	(1,107)	2.468	0.357	0.603	0.022
	DSW	4.87	(1,111)	16.436	<0.001***	>100	0.129
SSPT- DSWT	SSP	3.51	(1,113)	6.250	0.013*	3.205	0.052
	SN	1.34	(1,114)	1.237	0.536	0.336	0.010
	SSW	0.72	(1,116)	0.351	0.554	0.227	0.003
	DSP	-0.37	(1,115)	0.074	0.785	0.200	0.000
	DN	-0.50	(1,111)	0.193	0.661	0.215	0.001
	DSW	-3.15	(1,112)	7.286	0.016*	4.682	0.061

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected).

Simultaneous Spikiness Training vs Control

After training, the Simultaneous Spikiness Training group had greater accuracy relative to Controls on the Simultaneous Spikiness (p<0.001, BF_{10}>100, η_p^2=0.25) and Delayed Spikiness tasks (p=0.012 BF_{10}=7.30, η_p^2=0.06) when controlling for pre-training scores.

Delayed Switching Training vs Control

After training, the Delayed Switching Training group had greater accuracy relative to Controls on the Simultaneous Spikiness (p<0.004, BF_{10}=16.18, η_p^2=0.08), Delayed Spikiness (p=0.012, BF_{10}=8.83, η_p^2=0.07), and Delayed Switching tasks (p<0.001, BF_{10}>100, η_p^2=0.12) when controlling for pre-training scores.

Simultaneous Spikiness Training vs Delayed Switching Training

Both training groups made greater on-task gains relative to the other. The Simultaneous Spikiness Training group had greater accuracy relative to the Delayed Switching Training group on the Simultaneous Spikiness task after training (p=0.013, BF_{10}=3.20, η_p^2=0.05), when controlling for pre-training scores. Conversely, the Delayed

Switching Training group had greater accuracy relative to the Simultaneous Spikiness Training group on the Delayed Switching task after training (p=0.016, BF_{10}=4.68, η_p^2=0.06), when controlling for pre-training scores.

3.4.4 Transfer to components of the switching tasks

To further investigate whether there was any partial transfer to the switching tasks I analysed performance on the two judgment types within the switching tasks separately. This was primarily to examine whether practice on one judgment would improve performance on these judgments only in a switching context, or whether it would generalise to both judgment types. Summary statistics are provided in Table 3.4.

Table 3.4. Pairwise group ANCOVAs of post-training accuracy on the switching tasks by judgment type, adjusted for baseline performance

Task	Judgement	Post-training accuracy difference (%)	ANCOVA				
			df	*F*	*p*	BF_{10}	η_p^2
SSPT-Control							
SSW	Spikiness	3.97	(1,111)	6.02	0.048*	3.07	0.05
	Number	2.37	(1,111)	2.68	0.208	0.68	0.02
DSW	Spikiness	1.08	(1,110)	0.60	0.440	0.27	0.00
	Number	3.22	(1,110)	3.06	0.146	0.79	0.02
DSWT-Control							
SSW	Spikiness	2.34	(1,112)	2.09	0.302	0.55	0.01
	Number	4.00	(1,112)	6.33	0.039*	3.79	0.05
DSW	Spikiness	3.73	(1,111)	8.18	0.015*	8.19	0.06
	Number	6.33	(1,111)	11.27	0.003**	27.25	0.09
SSPT- DSWT							
SSW	Spikiness	1.73	(1,116)	1.35	0.302	0.35	0.01
	Number	-1.10	(1,116)	0.56	0.543	0.25	0.00
DSW	Spikiness	-2.75	(1,112)	3.59	0.122	0.99	0.03
	Number	-3.04	(1,112)	3.27	0.146	0.82	0.02

Note. Assessment task abbreviations: Simultaneous Switching (SSW); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected).

As expected, after training the Delayed Switching Training group had greater accuracy relative to Controls on the spikiness (p=0.015, BF_{10}=8.19, η_p^2=0.06) and enumeration judgments (p<0.01, BF_{10}=27.25, η_p^2=0.09) on the Delayed Switching Task, when controlling for pre-training scores. In addition, there was some evidence for partial transfer to the Simultaneous Switching Task, as the Delayed Switching Training group had greater accuracy relative to Controls on the enumeration judgments after training (p=0.039, BF_{10}=3.79, η_p^2=0.05), when controlling for pre-training scores. Finally, there was some evidence that the Simultaneous Spikiness Training partially transferred to spikiness

judgments on the Simultaneous Switching Task relative to Controls after training (p=0.048, BF_{10}=3.07, η_p^2=0.05), when controlling for pre-training scores. In other words, if people trained on the Delayed Switching Task then they improved on the enumeration half of the Simultaneous Switching Task relative to controls. Further, if people trained on the Simultaneous Spikiness Task they improved on the spikiness half of the Simultaneous Switching Task.

3.4.5 Task relationships and transfer

I tested whether the pattern of transfer could be predicted on the basis of various task relationships. There were three different ways of operationalising task relationships: i) the overall number of shared features, proportional to the total number of features, ii) whether the 'spikiness' feature was shared (because this was the only feature shared between the training groups) and iii) task correlations at baseline (see feature coding in Tables B.1 and B.2, and task correlation structure in Figure 3.7). I did this because I wanted to look at graded patterns of transfer across tasks, rather than a binary criterion of whether individual tasks show significant transfer or not. The former is likely to be more informative as to the nature of transfer.

For this analysis, I calculated each subject's individual task improvement, relative to the mean performance change for the control group (transfer). I then calculate how much of the variability in task improvement was explained by the three ways of operationalising task relationships. In other words, how well can transfer be predicted for a given task by the strength of its relationship with the respective training task?

I first z-scored (across groups) all of the post-training scores within each task, then fit a simple regression model for the Simultaneous Spikiness Training and Delayed Switching Training groups separately, wherein the difference in performance at post-training relative to the control group was the outcome variable, and degree of overlap (operationalised in three ways) was the predictor variable. This gave one beta co-efficient for each individual (i.e. how much each individual subject's pattern of transfer was determined by each of the three ways of calculating task relationship), and thus a distribution of beta coefficients across subjects (from which I derived Bayes factors). I then used 2000 bootstrapped samples to produce p-values (subsequently holm-corrected).

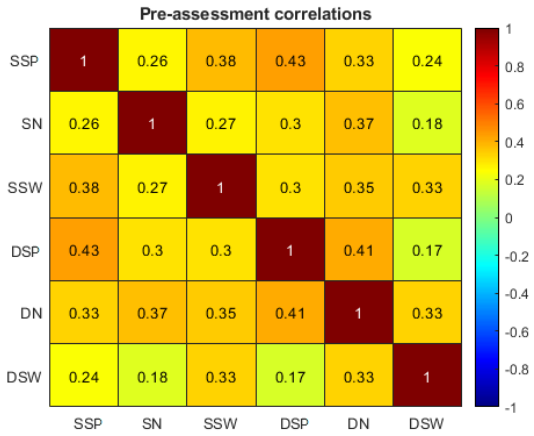
The proportion of shared features was significantly predictive for the Simultaneous Spikiness Training group (mean β=0.17, p=0.010, BF_{10}=5.52) but not the Delayed Switching

Training group (mean β =0.04, p=0.215, BF_{10}=0.20). This was repeated, but with the binary predictor of whether the tasks shared the spikiness feature. The results show that the spikiness feature was predictive of transfer for the Delayed Switching Training group (mean β=0.29, p<0.001, BF_{10}=228.85) and the Simultaneous Spikiness Training group (mean β=0.12, p=0.042, BF_{10}=1.29). Finally, I repeated this procedure once more but with the pre-training correlation values as predictors. The results showed that the correlations between training and assessment tasks at pre-training were not predictive of transfer across tasks for the Simultaneous Spikiness group (mean β=0.08, p=0.121, BF_{10}=0.50) or the Delayed Switching Training group (mean β=0.00, p=0.458, BF_{10}=0.15).

3.4.6 Correlations pre- and post-training

To further explore the relationships between tasks and how these might change as a function of training, I examined the correlations at pre assessment, post assessment, and the difference between the two (Figures 3.7 and 3.8). This mirrors the analysis conducted in the previous chapter. To establish whether correlations changed significantly following training I used a permutation method wherein I randomly sampled (n=2000) the pre and post assessment task performances for each group and calculated the pairwise changes in the correlation coefficients to estimate a distribution and produce p-values.

Figure 3.7. Pre-assessment correlations across groups.



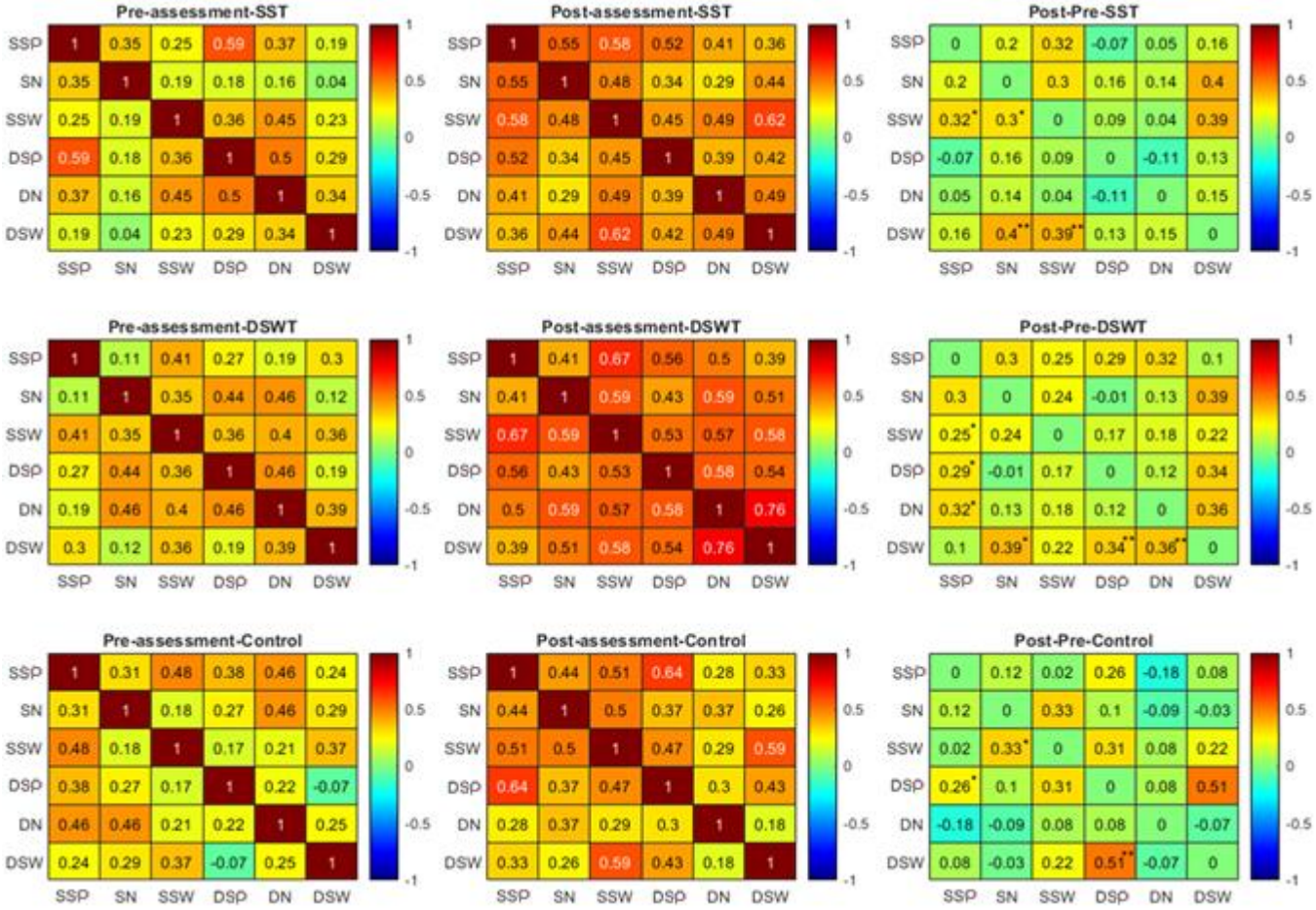
Note. Pre-assessment correlations across groups. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW).

For the Simultaneous Spikiness Training group, the following task pairs became more correlated with training: Simultaneous Switching-Simultaneous Spikiness (r-pre=0.25, r-post=0.58, p=0.042), Simultaneous Switching-Simultaneous Number (r-pre=0.19, r-

post=0.48, p=0.045), Delayed Switching-Simultaneous Number (r-pre=0.04, r-post=0.44, p=0.008), and Delayed Switching- Simultaneous Switching (r-pre=0.23, r-post=0.62, p=0.002).

For the Delayed Switching Training group, the following task pairs became more correlated: Simultaneous Spikiness-Simultaneous Switching (r-pre=0.41, r-post=0.67, p=0.034), Simultaneous Spikiness-Delayed Spikiness (r-pre=0.27, r-post=56, p=0.036), Simultaneous Spikiness-Delayed Number (r-pre=0.19, r-post=0.50, p=0.041), Delayed Switching-Simultaneous Number (r-pre=0.12, r-post=0.51, p=0.017), Delayed Switching-Delayed Spikiness (r-pre=0.19, r-post=0.54, p=0.006), and Delayed Switching-Delayed Number (r-pre=0.39, r-post=0.76, p<0.001).

Figure 3.8. Group correlations at Pre-assessment, Post-assessment, and the differences.



Note. Permutation sampling (n=2000) was used to form a distribution of the differences, p-values were derived by calculating the proportion of values greater than zero. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT).

For the Control group, the following task pairs became more correlated: Delayed Spikiness-Simultaneous Spikiness (r-pre=0.38, r-post=0.64, p=0.035), Simultaneous

Switching-Simultaneous Number (r-pre=0.18, r-post=0.50, p=0.040), Delayed Switching-Delayed Spikiness (r-pre=-0.07, r-post=0.43, p=0.002).

A number of task pairs changed in their strength of association following training for all groups. These were all in the positive direction, that is, the task pairs shared more common variance with one another following training. Whilst between group comparisons were not conducted, on the surface, the pattern of task pairs that change in association appear to be at least somewhat unique to each of the training groups and those that are shared seem to vary in magnitude. However, between group comparisons will need to be conducted to verify this.

3.5 Discussion

From the previous chapter we already knew that task relationships can change as a result of cognitive training, and that different patterns of transfer exist between individuals. Here, I evaluated how practice-dependent transfer is related to shared task features using a tightly controlled randomised design with a relatively large sample and adaptive control group. All of the tasks required same-different judgements on a common set of spikey shapes. The task components required to perform each task varied systematically, such that they formed a nested hierarchy. Training was then performed on two of the tasks: one was relatively ‘low’ in the hierarchy requiring just simultaneous judgments of shapes’ spikiness, whereas the other was relatively ‘high’, requiring delayed judgments of shapes’ spikiness or number of spikes in a switching paradigm. Using the full complement of tasks before and after training I could then test whether and how the effects of training on these ‘low’ and ‘high’ tasks cascade through their hierarchy. I asked the following questions: 1) Do participants make substantial on-task training gains? 2) Do the different training tasks generate different transfer patterns? 3) Are these transfer patterns predicted by the proportion of overlapping features? 4) Do some shared features contribute more to the transfer than others? 5) Is the direction of transfer unidirectional or bidirectional relative to the position of the training task?

Both training groups showed on-task improvements as well as selective transfer to other tasks, relative to active controls. Specifically, Simultaneous-Spikiness training transferred to a delayed-presentation variant but not to tasks requiring an enumeration judgement nor those requiring switching between judgements. The Delayed-Switching training transferred to two tasks requiring spikiness judgements but not to tasks requiring an

enumeration judgment nor to the other switching task variant with a simultaneous presentation type. In short, there was evidence of transfer to other tasks requiring the same basic spikiness judgement, but no evidence of transfer of 'switching' ability, or transfer to the enumeration judgement.

I also directly assessed whether task relationships defined in multiple ways could predict transfer patterns within the hierarchy. For both training groups, relative to the control, whether or not an assessment task required the spikiness judgement was significantly predictive of the pattern of transfer. For the Simultaneous-Spikiness training group, the overall overlap in features between the training and assessment tasks also significantly predicted the pattern of transfer, but not for the Delayed-Switching training group. Pre-training between-task correlations were not predictive of the pattern of transfer for either group. Moreover, as in Chapter 2, an exploratory analysis indicated that several between-task correlations changed following training.

3.5.1 Multi-component training resulted in broader transfer

The higher-level delayed switching training transferred to lower-level tasks requiring spikiness judgments, but the reverse is not true. Simple spikiness training did not transfer *up* the hierarchy to either of the tasks requiring switching. More precisely, these findings suggest that switching was a boundary condition for transfer within this task hierarchy. Furthermore, training on both judgments in the switching paradigm did not prevent transfer to the spikiness tasks despite this training group receiving half as much practice on these judgment types. That the Simultaneous-Spikiness training did not transfer to either of the switching tasks, suggests that the additional demands imposed by switching are enough to nullify the Spikiness judgment training effects. That is, getting better at one of the constituent tasks does not influence the ease with which participants can switch between the tasks.

As previously discussed in the introduction, the idea of a task-set or cognitive routine may help explain these findings. Simple spikiness training may engender a task-set/cognitive-routine that serves to bias information processing and prevent interference from irrelevant stimulus information (e.g. number of spikes) and/or response-mappings, helping to maintain an improved on-task performance (Rogers & Monsell, 1995; Dreisbach & Wenke, 2011; Gathercole et al., 2019). However, this same task-set/cognitive-routine may be exogenously activated in the context of the switching tasks due to the use of the same stimuli and similar demands across tasks. This could cause some initial negative transfer effects (Rogers &

Monsell, 1995; Dreisbach & Wenke, 2011) and may explain why switching was a boundary condition here. Conversely, switching training may engender a more relaxed task-set/cognitive-routine leading to broader transfer effects on novel untrained tasks (Dreisbach & Wenke, 2011; Sabah et al., 2019; Gathercole et al., 2019).

3.5.2 Task relationships have mixed predictive power for transfer

Establishing taxonomic relationships amongst cognitive tasks, and thus the overlap between them, is required for making quantitative predictions about the magnitude and distance of transfer, both of which are important for theoretical progress (Reder & Klatzky, 1994; Barnett & Ceci, 2002; Taatgen, 2013; Gathercole et al., 2019; Smid et al., 2020). Despite this, few training studies explicitly quantify relationships between training and assessment tasks such that they can be used in predictive models of transfer, though there exist some notable exceptions (Singley & Anderson 1985; Taatgen, 2013; Gathercole et al., 2019). In this study I explored three simple measures of relatedness between each of the trained tasks and the remaining untrained assessment tasks to predict transfer. I found that the proportion of shared features between the training and assessment task was predictive of transfer in the case of the Simultaneous Spikiness group but not the Delayed Switching group. Whereas the binary measure of whether the spikiness features was shared was predictive of transfer for both groups. Pre-training correlations were not predictive of transfer for either training group. In short, transfer was most consistently predicted by the presence of a specific shared feature (spikiness judgement) rather than the more general measures of relatedness (number of shared features or correlations) across different training paradigms.

The lack of predictive power for pre-training correlations suggests that two tasks may be predictive of one another prior to training, but this does not mean that an improvement on one will transfer to the other. Presumably, this is because two tasks may share key cognitive processes but, as discussed in Chapter 2, the processes recruited likely change or are re-weighted as a function of training. This re-weighting of task relationships after training is also reflected in the changing correlational strengths between tasks. Two tasks may share many of the same cognitive processes both before and after training but unless the ‘key ingredient’ acquired during training, i.e. the process that is responsible for the improvement, is also applicable to the untrained task then we will not see transfer (Gathercole et al., 2019; Taatgen, 2013).

Taken together, these findings suggest that the patterns of transfer in this study do not manifest in neat accordance with any of these simple measures of relatedness. Instead, they seem to further echo the sentiment of prior research that not only is transfer tied to specific features but also to the specific *task-context* in which these features arise (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Sala & Gobet, 2019; Soveri et al., 2017).

3.5.3 Task relationships change following training

Following training several correlations between pairs of assessment tasks substantially increased in each group. This echoes the findings from Chapter 2, in which task pairs (as represented by a SOM) also changed in their correlational strength following training. By some definitions of task similarity, the more common variance two tasks share, the more similar they are. In contrast, definitions of task similarity based on feature relationships remain constant over time. A change in correlational strength between two tasks following training is suggestive of a change in the cognitive processes recruited. Whilst in Chapter 2 we observed both positive and negative changes in strength, here we observed only positive changes. This perhaps reflects the fact that the tasks used in the study were all close relatives of one another by design, whereas the tasks used in Chapter 2 were more diverse (although these were not explicitly defined). It may be that the range of processes recruited becomes more similar with training or that certain key processes are more heavily weighted across tasks. It is interesting that between-task correlations are subject to change following training, to speculate that these may be indicative of a change in the processes recruited, and that these may differ depending upon the type of training. However, without a more formal and rigorous investigation, including between group contrasts, it is hard to interpret these findings any further. Nonetheless, this may be a fruitful future avenue of research. Moreover, the fact that these relationships are subject to change following training should be considered in future research utilising correlational analyses.

3.5.4 Switching does not transfer across presentation types

Previous studies have shown that task-switch training consistently transfers to other similarly structured switching tasks (requiring different categorical judgements about objects in pictures), evidenced primarily by reduced reaction time switch costs, thought to reflect a reduced interference caused by switching demands (Dorrenbacher et al., 2014; Karbach & Kray, 2009). Despite substantial on-task gains, the Delayed-Switching training failed to transfer to the Simultaneous-Switching paradigm in terms of accuracy or reaction time. Thus,

presentation type was a boundary condition for transfer suggesting that the switching skills acquired during training are tied specifically the delayed mode of presentation. It is unclear why this might be, however one key difference between this study and others before is the emphasis placed on accuracy by instructing participants to be as accurate as possible rather than the more commonplace instruction to be as fast and as accurate as possible. Previous research also failed to find transfer with respect to accuracy (Dorrenbacher et al., 2014; Karbach & Kray, 2009), therefore it is possible that this effect is specific to reaction time, something that was not encouraged by the task-switch training in this study.

I further investigated transfer effects after splitting switching task performance into its constituent spikiness and enumeration judgments. There was a small amount of evidence for transfer of the Delayed-Switching training to the enumeration judgement type in the Simultaneous-Switching task. This suggests that whilst the transfer of skills pertaining to the spikiness judgment (in a switching context) were bounded by presentation type those pertaining to the enumeration judgement were not (in a switching context). In addition, there was anecdotal evidence that Simultaneous-Spikiness training partially transferred to spikiness judgments on the Simultaneous-Switching task. This suggests a graded pattern of transfer, whereby practice on one perceptual judgment may transfer to trials of the same type in a switching context, but only when the presentation type is consistent (i.e. simultaneous).

3.5.5 Transfer was constrained by the type of perceptual judgment

The generalisability of training gains appeared to be bounded by the judgement type in both groups, as neither showed substantive transfer to tasks involving enumeration judgements. This is most clearly demonstrated by the fact that Simultaneous-Spikiness training transferred to a task (Delayed-Spikiness) comprised of an identical judgement but different presentation, and conversely did not transfer to a task (Simultaneous-Number) comprised of a different judgement but identical presentation. Moreover, the transfer effect for Simultaneous-Spikiness training to the Delayed-Spikiness task was small-medium, whereas the on-task effect was large. This suggests that the cognitive processes acquired during training were specific to both the spikiness judgement and simultaneous presentation mode in tandem (Ahissar & Hochstein, 2004; Dosher & Lu, 2017).

One possible explanation for this specificity is that participants are learning to better represent the spikiness feature by reducing the signal to noise ratio of population codes in the visual cortices via the updating of synaptic weights (Dosher & Lu, 2017; Fahle, 2005). This

may reduce ambiguity when making spikiness judgements but not enumeration judgements, due to the number of spikes feature being differentially encoded and relatively unaffected by training. Relatedly, there may be alterations to attentional or executive processes responsible for orchestrating the parsing of spikiness representations and subsequent decision action mappings (Ahissar & Hochstein, 2004; Dosher & Lu, 2017; Taatgen, 2013). Finally, the cognitive routine framework of transfer (Gathercole et al., 2019) emphasises how novelty necessitates the acquisition of new cognitive routines that engender transfer when they can be applied to similarly structured tasks. It is plausible that the routines used for making rapid enumeration judgements are relatively well established prior to the study and thus show less room for improvement and transfer.

3.5.6 Limitations

The current study had several limitations. It could have benefited from a fuller range of training groups. For example, including a group that trained solely on the enumeration judgement would help verify whether this judgment type was potentially capped due to prior experience/lack of sensitivity or whether it was the case that more training was required for improvements to manifest. Similarly, including a group that trained on the Delayed-Spikiness task would help determine the extent to which the observed transfer to and from this task was limited due to the presentation type. Another limitation is that participants received only a very small amount of training (three sessions per group) relative to most other training studies, so we cannot know if these findings would extend to longer periods of training which may have induced wider patterns of transfer. This decision was made because pilot data suggested that on these very simple tasks, improvements were rapid (in accordance with the power law of practice; Newell & Rosenbloom, 1981), and I wanted to maximise the sample size. Nonetheless, it may be that transfer would be more extensive if more extended periods of training were given.

3.5.7 Conclusion

In summary, training at different levels within a feature based taxonomic task hierarchy produces different transfer patterns. The design allowed different types of task overlap to be quantified. The best predictor of whether transfer would occur was whether the tasks shared a particular feature – the spikiness judgement. However, for one training group transfer was also related to the overall proportion of shared features. Finally, whether task performance is correlated pre-training is not a good predictor of transfer patterns.

Collectively, these findings provide a further demonstration of the specificity of transfer and provide an experimental exploration of the nature of task overlap that is crucial for the transfer of performance improvements.

Chapter 4: Training and transfer within nested tasks: a change detection training paradigm.

4.1 Introduction

Recent studies highlight the prevalence of feature-specific improvements that occur following typical cognitive training paradigms (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019). Using tasks in both the assessment and training phases that are nested and vary systematically with respect to specific external task-features, is one approach toward better understanding the nature of this specificity and its boundary conditions (Holmes et al., 2019; Norris et al., 2019). This was the approach I took in the previous chapter, revealing that the best predictor of transfer was a specific shared feature, common to both the training and assessment tasks. However, the tasks used in the previous chapter were relatively low level (e.g. discriminating between features of two shapes). It is possible that task complexity influences the patterns of transfer (Taatgen, 2013; Assihar & Hochstien, 2004; Dosher & Lu, 2017; Dreisbach & Wenke, 2011) – where the nested tasks recruit higher-level cognitive processes, the patterns of transfer may differ.

The change detection paradigm is a popular measure of visual working memory (VWM) that lends itself nicely to a nested approach for both practical and theoretical reasons. In this chapter I use nested change detect task (CDT) variants to investigate which skills are acquired during training and whether these are feature-specific or generalise more broadly. There exists a rich theoretical backdrop pertaining to both the CDT tasks themselves and the training context. Whilst much of the motivation for the approach in this chapter is shared with that of the previous, I will begin with a brief re-cap of the theoretical issues that motivate the empirical work before introducing the current training paradigm and its potential implications.

4.1.1 Motivational re-cap

Due to the far-reaching implications for education and wellbeing, multiple studies in recent decades have tested training induced transfer effects (Green & Bavelier, 2008; Sala & Gobet, 2019; Simons et al., 2016). Or in other words, when does practice on one task carry over to another? As reviewed extensively earlier in the thesis, there is a plenty of evidence for near transfer, but far transfer remains controversial (Green & Bavelier, 2008; Au et al., 2014; Klingberg, 2010; Jaeggi et al., 2008; Holmes et al., 2009 Melby-Lervag et al., 2016; Sala & Gobet, 2019; Simons et al., 2016; Soveri et al., 2017; Gathercole et al., 2019; Smid et al.,

2020). That is, training improvements often carry over to similar tasks within domain, but rarely translate to improvements to more distant tasks. Moreover, even within domain improvements appear bound to specific task features and may even fail to transfer between two tasks that differ by only a single feature (Gathercole et al., 2019).

There is considerable interest in theories that explain when transfer will, and will not, occur. These all focus on operationalising task similarity, which, as has already been emphasised, is a non-trivial and fundamental issue in the field of cognitive training and cognitive science more generally (Barnett & Ceci, 2002; Kievit et al., 2011; Maul et al., 2016; Smid et al., 2020; Meyer et al., 2001; Miyake et al., 2000; Taatgen, 2013). To summarise briefly, there are some key ways of defining task similarity: One approach is to define the similarity between tasks according their correlational properties (behaviourally or physiologically). Although, as was demonstrated in the previous two empirical chapters, neither the correlations themselves, nor latent constructs, are predictive of transfer, and are themselves subject to change as a function of training. A second approach is to organise tasks according to the composition of hypothetical processing components (Anderson, 1982; Feldman & Ballard, 1982; Singley & Anderson, 1985; Taatgen, 2013; Yang et al,. 2019). However, this requires many assumptions and extensive practical, theoretical, and technical experience and knowledge, rarely agreed upon and not readily available for many researchers. A third approach is to define task similarity through the identification, specification, and variation of the extrinsic task-features from which tasks are comprised (Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Soveri et al., 2017). Such task analytical approaches are stable and relatively straightforward. However, they vary in resolution and still require mapping onto theory, both of which can dramatically affect the interpretation of any training outcomes.

Again, as previously stressed, these three approaches (i.e. correlational, hypothetical modelling, task analytical) are not mutually exclusive, likely map onto one another, and should be used in tandem when possible. A goal of this thesis has been to make systematic and tightly controlled manipulations of extrinsic task features to establish the boundary conditions of transfer and inform cognitive theory (Katz et al., 2018; Redick, 2019; Sala & Gobet, 2019; Von Bastian & Oberauer, 2014; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Smid et al., 2020; Taatgen, 2013). In this vein, the current chapter represents the next step in this process, using a change detection task (CDT) paradigm due to its potential for subtle feature manipulations and a rich theoretical backdrop.

4.1.2 Visual working memory and the change detection task

VWM can be conceptualised as the limited capacity ability to maintain and manipulate visual information over short periods of time in service of ongoing task demands (Baddeley & Hitch, 1974; Barak & Tsodyks, 2014; Cowan et al., 2005; Christophel et al., 2017; Luck & Vogel, 2013). VWM ability is of great interest because it strongly correlates with measures of cognitive control, intelligence, and educational attainment, as well as being essential for successfully performing many everyday tasks (Alloway & Alloway, 2010; Cowan et al. 2005; Cowan et al. 2006; Johnson et al. 2013; Diamond, 2012; Fukuda et al. 2010; Luck & Vogel, 2013).

The change detection task (CDT) paradigm is a popular measure used to tap VWM (Buschkuehl et al., 2017; Luria & Vogel 2011). There are several variations of the CDT paradigm, each of which has provided useful insights in its own right. However, they all follow the same general form: a brief presentation of a memory array containing a set of visual stimuli, followed by a retention interval, and then a test array containing one or more probe stimuli. Participants are required to judge whether (or how) some aspect of a probe stimulus has changed relative to its counterpart (e.g. same location) in the memory array. Conceptually, these three phases correspond roughly to: encoding, maintenance, and retrieval (Luck & Vogel 1997; Buschkuehl et al., 2017; Fougnie & Alvarez , 2011).

A seminal study conducted by Luck & Vogel (1997) investigated VWM using a series of CDT paradigms. They found that participants had a high accuracy for detecting changes on a single feature dimension (e.g. colour) for up to four objects, with a sharp decline thereafter. Crucially, they also found that participants had an almost identical accuracy for detecting changes across multiple feature dimensions (e.g. colour and orientation). They took this as evidence to suggest that VWM stores discrete-object level representations rather than individual features and has a capacity of 3-4 items. Others too have proposed that VWM consists of a small number of slots, each of which stores a single integrated visual object, or chunk, with fixed precision (slot model; Cowan, 2001; Luria & Vogel, 2011; Vogel et al, 2001; Rouder et al., 2008; Zhang & Luck, 2008). However, subsequent lines of research have questioned both the proposed capacity estimates and the nature of VWM representations (Alvarez & Cavanagh, 2004; Bays et al., 2009; Schneegans & bays, 2016; Souza & Oberauer, 2016).

Continuous response paradigms allow researchers to examine the precision of recall. As the number of items in the memory array (set size) increases, response precision decreases exponentially, indicating that the resolution with which representations are stored trades-off against the number of items being stored (Alvarez & Cavanagh, 2004; Bays et al., 2009; Awh et al., 2007; Bays & Hussain, 2008; Wilken & Ma, 2004). These findings appear consistent with a model of VWM in which the precision of a stored representation depends upon the proportion of a common VWM resource allocated to it (resource model; Bays & Hussain, 2008; Bays et al., 2009). Participants do not know which item from the memory array will be probed, and will therefore share resources out amongst the items, hence why performance declines as the number of items in the memory array increases. Formalisations of the resource model successfully capture the behavioural data and the *appearance* of a capacity limit without necessarily imposing a fixed upper limit on the number of items stored in VWM (Schneegans & bays, 2016; Schneegans & bays, 2017). Moreover, errors under the resource model appear well accounted for by failures to accurately decode feature values from activity in neuronal populations, providing physiological plausibility to the model (Taylor & Bays, 2020; Schneegans & bays, 2017; Schneegans et al, 2020).

Importantly, CDT paradigms require memory not only for the to-be-reported dimension (e.g. colour or orientation) but also the cue dimension (e.g. location). This can sometimes lead to what are called ‘swap errors’, errors thought to occur when features of non-target items interfere with the recall of the target item, due to a failure in correctly integrating, or ‘binding’, information across the report and cue dimensions (Schneegans & bays, 2016; Schneegans & bays, 2017). Others have found that when participants are required to recall two cue features, such as colour and orientation, from objects in a memory array within a continuous response paradigm, the errors for each of the features are weakly correlated with one another and thus appear to be relatively independent (Bays et al., 2011; Fougnie & Alvarez , 2011). Relatedly, others have found a monotonic decrease in CDT performance as the complexity of objects in the memory array increases (Alvarez & Cavanagh, 2004; Wheeler & Treisman, 2002; Olson & Jiang, 2002; Xu, 2002). Together, the above findings imply that VWM resources are allocated, at least somewhat independently, to features within a visual scene rather than objects. Accordingly, alternative models have been proposed, in which the information required to combine features into integrated objects is maintained separately and independently from the information pertaining to the features

themselves, which in turn are maintained in at least a somewhat dissociable and limited fashion (Brady et al., 2011; Wheeler & Treisman, 2002; Fougnie & Alvarez , 2011).

Marshell & Bays (2013) argue that by attending to one feature, all others are automatically encoded but that maintenance of specific features can be modulated by attention. This sentiment is echoed by research looking at the well-established retro-cue effect. A retro-cue is an attentional cue presented during the maintenance phase of a trial, providing an indication as to which item will be probed at retrieval, and thereby orienting attention to the internal representation of that item (Griffin & Nobre, 2003; Souza & Oberauer, 2016). Retro-cues can substantially improve VWM performance in terms of both accuracy and reaction time (Astle et al., 2012; Griffin & Nobre, 2003; Heur & Schubo, 2016; Sligte et al., 2008; Sligte et al., 2010; Souza & Oberauer, 2016). This suggests that VWM capacity is greater during maintenance than no-cue trials alone would suggest but that it is negatively impacted by retrieval and/or response processes (fragile memory). It also provides a powerful demonstration of the important role that top-down attention plays in modulating the content of VWM (Griffin & Nobre, 2003; Heur & Schubo, 2016; Oberauer & Hein 2012; Souza & Oberauer, 2016). By allowing VWM resources to be flexibly allocated in time and space, visually salient items may be remembered with enhanced precision and protected from interference, whilst less important items may be weakened thereby freeing up resources for use elsewhere (Astle et al., 2012; Heur & Schubo, 2016; Souza & Oberauer, 2016; Marshell & Bays, 2013).

Whilst the capacity and nature of VWM are contentious, there is evidence to suggest that VWM resources may be flexibly allocated via attention. As such, training people on CDT paradigms may lead to changes in their attention to certain features, thereby making them more or less salient. By including several CDT training variants we are able to ask whether any changes in attention occur, and if so, are they feature specific or do they generalise more broadly? Moreover, including retro-cues on half of the assessment task trials, enables one to test whether these training effects impact upon the spatial allocation of attention during the maintenance phase, or alternatively whether they impact performance independently of this.

4.1.3 Change detection task training

Despite its popularity, only a handful of studies to date have looked at the effects of training on the CDT paradigm. This is perhaps due to a few early studies reporting only small

effects for on task improvements following practice (Eng et al., 2005; Olson & Jiang., 2004; Olson et al., 2005). However, these studies were primarily investigating the effect of long-term memory traces within sessions, rather than transfer effects and so were necessarily limited in scope. More recently, Xu et al (2017) found small to moderate improvements across 30 sessions of training on a CDT paradigm and found performance to be relatively stable over time with respect to both within and between subject performances. However, this training was non-adaptive. Adaptive training is thought to enhance performance above and beyond non-adaptive training (Buschkuchl et al., 2017; Jaeggi et al. 2014). Using an adaptive training regime Buschkuchl et al (2017) found substantial on-task improvements following 4 hours of colour-CDT training (20-25% increase in Cowan's K). They also found very little indication of transfer across several other measures, including a colour resolution task and a similarly structured CDT paradigm that used more complex stimuli. They suggest that the training likely had an impact on early stimulus representations specific to the task but not on the ability to retain details of the stimuli nor any higher-level cognitive processes. Moreover, the study did not contain a control group and thus it cannot control for any test-retest effects.

Only a few studies have used a control group and they found mixed results (Gaspar et al., 2013; Moriya, 2019; Norris et al., 2020). Gaspar et al (2013) trained participants adaptively on an object-CDT paradigm by progressively reducing the display time of the memory array, encouraging faster encoding. This training substantially improved on-task performance, but these improvements did not transfer to their 'near-transfer' measure of a similar object-CDT paradigm, nor to their 'far-transfer' measure of a flicker-CDT paradigm. The authors concluded that the training improvements were due to an increased familiarity with the training stimuli rather than an improvement in any change detection ability more broadly. However, as the authors note, their training targeted processes associated with faster encoding of the stimuli and so we do not know whether training aimed at other processes associated with capacity may have produced a different outcome. Moreover, despite labelling one of the assessment tasks as a 'near-transfer' measure, it still differed from the training task with respect to timings, masking, stimulus categories, and set size manipulations, thus making it difficult to establish precisely what constrained transfer.

Alternatively, Norris et al (2020) trained participants adaptively on a colour-CDT paradigm by progressively increasing the number of items in the memory array (set size), encouraging enhancements in capacity. This training substantially improved on-task performance and these improvements transferred to an untrained orientation-CDT paradigm

that was identical to the training task except the memory items were oriented bars instead of coloured squares. This suggests that at least some of the cognitive processes acquired during colour-CDT training are generalizable, such that they can be utilised across stimulus type. In contrast, Adam & Vogel (2018) found feature specific transfer after training participants on a colour-whole-report task paradigm. Specifically, they found that colour-whole-report training transferred to a colour-CDT paradigm (same feature modality, different response requirement) but not to an orientation-whole report task (different feature modality, same response requirement). However, the training used here was not adaptive. Moreover, the response requirements of whole report tasks are a lot more involved compared with those in the CDT and so any transfer benefits occurring at encoding or maintenance may have been 'washed out' due to the interference cause by the response requirements.

Another recent study by Moriya (2019) trained participants on two highly similar orientation-CDT paradigms and tested participants on equivalent assessment versions of these tasks before and after training. One task was considered a 'quantity task' and the other as a 'quality task'. The quantity task contained set sizes of 4, 6, and 8, with probe offsets of 45 degrees, whilst the quality task contained set sizes of 2, 4, and 6, with probe offsets of 15 degrees. As in Norris et al (2020), participants trained adaptively by progressively increasing the set size of the memory array. Moriya found that training on the quantity version of the task transferred to the quality version but not the other way round. Moriya interpreted this finding as support for the idea that training enhances the allocation of limited VWM resources for both the quantity and quality of the memory items, and that the two share overlapping processes. However, if this were so, it is unclear as to why we would see transfer one way but not the other (although it was trending in that direction and may simply be a power issue). Moreover, the nature of the quality training was adaptive for set size and not offset, thus it could be argued that it was also training quantity but just at smaller offsets. An alternative interpretation of these results is that the quality task is a watered down version of the quantity task, with higher variance within set size due to the increased difficulty provided by the smaller offset. Nonetheless, the positive transfer found in this study is further evidence that skills acquired during CDT training may generalize to similar contexts. However, given that offset difficulty was not varied systematically at assessment we do not know whether the transfer effect found here represents an improvement in perceptual accuracy, an increase in the ability to store more objects in memory, or both (Bays & Hussain, 2008; Bays et al., 2009).

The CDT appears to be underexplored paradigm in the context of training and thus the boundary conditions for transfer within this otherwise popular working memory paradigm remain unclear. Given the substantial on task training gains following adaptive set size training (Buschkuchl et al., 2017; Norris et al., 2020; Moriya, 2019), the potential of transfer to similarly structured variants (Norris et al., 2020; Moriya, 2019), and the unresolved mechanisms of improvement (e.g. quantity vs quality), the current training study appears promising and well situated to further inform these areas.

4.1.4 Overview

There is increasingly convergent evidence to suggest that the transfer engendered by typical training studies is tied to specific task features – as I showed in the previous chapter (see also: Melby-Lervag et al., 2016; Sala & Gobet, 2019; Simons et al., 2016; Gathercole et al., 2019; Norris et al., 2019). The change detection task (CDT) paradigm lends itself well to subtle feature manipulations, which have proved fruitful in furthering our understanding of visual working memory (VWM; Luck & Vogel 1997; Luck & Vogel, 2013; Luria & Vogel 2011; Alvarez & Cavanagh, 2004; Bays et al., 2009; Schneegans & bays, 2016; Souza & Oberauer, 2016). Whilst the nature of VWM and its representations remain contentious, there is evidence to suggest that VWM resources can be flexibly allocated via attention to aid task performance (Griffin & Nobre, 2003; Luck & Vogel, 2013; Schneegans & bays, 2017; Marshell & Bays, 2013; Souza & Oberauer, 2016; Heur & Schubo, 2016). However, due to the limited number of randomised controlled training studies using the CDT paradigm, it remains unclear just how malleable the VWM resources used to perform the CDT are, or the extent to which changes that occur as a result of experience are tied to specific features or generalise more broadly. Given the potential for on-task training gains (Buschkuchl et al., 2017; Xu et al., 2017), transfer between similarly structured variants (Moriya, 2019; Norris et al., 2020), as well as a rich theoretical backdrop, the CDT paradigm provides a good training context for exploring the relationship between task overlap and training transfer.

4.1.5 The present study

The present study investigated which skills are acquired during adaptive training on a set of three hierarchically nested change detection tasks (CDTs) and the potential boundary conditions for the transfer of these skills. To do so, I conducted a large online training study, powered for small-medium effect sizes. I used three CDTs that were structurally almost identical but varied subtly from one another with respect to their specific judgement

requirements, I also used a digit span task as an active control condition. There were assessment versions of the four tasks and each also had a training version counterpart. Participants completed twelve sessions of training, which were adaptive with respect to the set-size of the memory array (span length in the digit span training). In all three CDT tasks, participants were presented with a memory array containing a number of arrows of various colours and orientations, followed by a retention interval, and then a test array containing a single probe stimulus. Importantly, the probe stimulus was always offset relative to its counterpart (same locations) in the memory array, with respect to both its colour and orientation, by either a small, medium, or large degree. In each of the three CDT tasks participants were required to judge the circular direction (clockwise or anti-clockwise) of change for either the colour, orientation, or both the colour and orientation of the arrows, respectively. On the CDT assessment tasks, half of the trials included a retro-cue in the second half of retention interval, cueing participants toward the location of the stimulus to be tested.

This design was motivated by, and allowed me to ask, the following questions:

1. Does training lead to the acquisition of skills that enhance the number of items stored in memory, the precision of those items, or both?
2. Does training affect the allocation of spatial attention during VWM maintenance?
3. Are the skills acquired during training in the single judgement tasks bound to their specific judgement types of colour and orientation, or do they transfer to one another?
4. Do the skills acquired during training in the single task conditions transfer ‘up’ the task hierarchy to a dual judgement (Orientation and Colour) task and vice versa?

I pre-registered the study (https://osf.io/nb5m7), and in spite of mixed evidence, made the corresponding predictions:

1. Given evidence for the modulation of both the quantity (Buschkuchl et al., 2017; Griffin & Nobre, 2003; Norris et al., 2020; Moriya, 2019; Heur & Schubo, 2016) and quality (Heur & Schubo, 2016; Bays et al., 2009; ; Marshell & Bays, 2013; Moriya, 2019) of VWM representations via spatial attention and experience I predicted that the training would enhance on-task performance in terms of both the number of items held in memory and the precision of the representations of those items. However, I acknowledged that these two factors interact with one another and so this may be hard to tease apart (Alvarez & Cavanagh, 2004; Bays et al., 2009; Bays & Hussain, 2008).

2. I speculatively predicted that training gains will have a greater impact on processes at the encoding/early maintenance phase of the trial due to the relatively long presentation times used here, leaving more room for strategically improving encoding efficiency, especially at higher set sizes (Gaspar et al., 2013; Vogel at el., 2006), and evidence suggesting an increased value of attentional modulation at encoding/early maintenance phases of a trial versus later on (Griffin & Nobre, 2003; Astle et al., 2012). If this is so, then we should not expect training to interact with the presence of the retro-cue during the post-encoding maintenance phase. Alternatively, if extended CDT practice essentially trains participants to allocate spatial attention strategically during the maintenance phase, then training should interact with the retro-cue / no cue manipulation. When there is a retro-cue everyone allocates their attention strategically, regardless of the training condition.
3. If maintaining representations with increased precision requires more focused attention (Alvarez & Cavanagh, 2004; Bays et al., 2009; Bays & Hussain, 2008; Oberauer & Hein, 2012) and memory for within object features can fail independently (Bays et al., 2011; Fougnie & Alvarez , 2011), then training in the single feature conditions may encourage a greater saliency for specific features, and thus I predicted that any training gains <u>for quality</u> of the representations will not transfer across judgement types. On the other hand, given that maintaining a greater number of items appears to require a broader focus of attention (Alvarez & Cavanagh, 2004; Bays et al., 2009; Bays & Hussain, 2008; Oberauer & Hein, 2012), wherein features of the objects appear to be bound to a certain extent (Luck & Vogel, 2013; Luria & Vogel 2011), and the prior evidence for the transfer of quantity gains across judgement types (Norris et al., 2020), I predicted that training gains <u>in quantity</u> will generalise across judgement types.

I did not make any specific predictions at pre-registration about the vertical direction of transfer specifically. However, it would seem reasonable, based on the findings from the previous chapter to suggest that a varied exposure of features at a higher level may encourage broader transfer, and thus that dual judgement training would transfer to both of the lower-level tasks, whilst single feature conditions may be constrained more specifically and not transfer ‘up’.

NB. Unfortunately, due to technical problems getting data from the server, what follows involves slightly less data than intended. In essence, the server interface struggles to handle the data files because they are so large. We are working to remedy this, but it will take some time. We currently have assessment data (i.e. the pre and post-training data) from 90% of our participants. However, the on-task training data cannot be extracted at present. The on-task data themselves are unlikely to be very interesting, because we are simply expecting substantial gains across all four groups. Moreover, we have assessment versions of all the training tasks so can check. Nonetheless, hopefully, I will be able to gain full access to the data in the near future and carry out an analysis with greater power and with the on-task training results.

4.2 Materials and methods

4.2.1 Ethical approval

This study received ethical approval from the Cambridge Psychology ethics committee, University of Cambridge, application number: PRE.2019.046. All participants provided informed consent by checking a box to confirm they had fully understood the implications of participation and their right to withdraw.

4.2.2 Participants

Unfortunately, at the time of writing there were issues retrieving the full data from the JATOS server and as such I was not able to include all the participants or include full demographic information

The final sample herein (see 'Data exclusion'), consisted of 168 English speaking adults with normal/corrected vision aged between 18 and 35 years of age. Participants were recruited via 'Prolific', a platform for recruiting and paying people to participant in online experiments. Participants were paid at a rate of £6 per hour and received an £8 bonus upon the satisfactory (i.e. not suspect of low effort or cheating) completion of all sessions. Participants were randomly assigned to one of the four training conditions: Orientation-CDT (N = 42), Colour-CDT (N= 41), Dual-CDT (N=41), and Digit-Span (control, N=44). All participants completed two assessment sessions, one before (pre), and one after (post), completing 12 sessions of adaptive training.

4.2.3 Assessment tasks and procedure

In each assessment phase (pre- and post-training) participants completed all four assessment tasks: (1) Colour-Change-Detection-Task (Col-CDT); (2) Orientation-Change-Detection-Task (Col-CDT); (3) Dual-Change-Detection-Task (Dual-CDT), split into two response types of Orientation (Dual-Ori-CDT) and Colour (Dual-Col-CDT); and (4) Digit-Span-Task (DS). Each assessment phase took approximately 90mins to complete. In both sessions, participants first completed the single-response Colour-Change-Detection and Orientation-Change-Detection tasks in a randomised order, followed by the dual-response (Colour & Orientation) Dual-Change-Detection-Task, and finally the Digit-Span task. All tasks were coded using JavaScript (jsPsych; De Leeuw, 2015), HTML, and CSS in house. Provided below are detailed written accounts of each of the task materials and procedures. Also see figure 4.1 for a graphical depiction of the CDT task trials.

Change Detection Tasks (CDTs)

All three CDTs required participants to make two-alternative forced choices (2AFC) about the direction in which a probe-stimulus had changed relative to its counterpart in the memory array presented prior to a retention interval. In each, participants were required to make these judgements according to changes in orientation, colour, or both. Participants were instructed to press the 'J' key when making a clockwise-response or the 'F' key when making an anticlockwise-response. They were instructed to be both accurate and fast. Participants received explicit step-by-step instructions along with examples and a small number of practice trials for each of the tasks. Attention checks were in place following each set of task instructions to help ensure that participants understood what was required; these were in the form of multiple choice arrays containing each of the judgement requirements, participants had to choose the appropriate judgement to continue with the task, a failure to do so resulted in a repetition of the task instructions. Participants were provided with eight simple practice trials, a failure to score above a certain accuracy threshold (70% and 60% for the single-response and dual-response tasks respectively) resulted in a single repetition of the practice trials. Again, this was to help ensure participants understood the requirements and response mappings before starting the task. I considered this especially important in this study due to its online nature and the high similarity between tasks. Participants completed two blocks on each of the three CDT tasks, with a short break in between each block. Participants received immediate feedback after each trial and statistics about their overall performance (mean accuracy and RT) at the end of each task.

All the CDT assessment tasks followed the same general form in terms of their procedural structure, timings, and the stimuli presented, the order and content of the phases in any given trial was as follows: fixation (600ms), memory array (600ms), retention interval (1000ms), test array (maximum 5000ms per response), feedback (300ms), inter trial interval (500ms). The details of the stimuli presented in each of these phases is outlined below.

Fixation

A central fixation cross with a font size of 100px is presented for 600ms within a 600px by 600px grid square.

Memory array

An evenly spaced 4x4 grid square (grid-height: 600px, grid-width: 600px; cell-height: 150px, cell-width: 150px) containing either 2, 4, or 8 coloured arrows was presented for 600ms. Each arrow was 120px in length with a line-width of 13px and an arrow-head-width of 21px. Each arrow was positioned centrally within a random cell, orientated randomly about a circular space divided evenly into increments of 5 angular degrees (360/5 = 72 potential starting orientations), and coloured randomly about a Hue Saturation Lightness (HSL) geometric colour space, with hue mapping to a circular space and divided evenly into increments of 5 angular degrees (360/5 = 72 potential starting hues), whilst saturation and lightness were held constant at 100% and 50% respectively.

Retention interval

Trials were divided evenly into retro-cue and no-cue trials. On no cue trials a blank 600px by 600px grid square was presented for 1000ms. On retro-cue trials a blank 600px by 600px grid square was presented for 500ms, followed by a 600px by 600px grid square presented for 500ms in which one of the grid cells was outlined, cueing the participant to the relevant target location for the upcoming probe-stimulus.

Test array

On each trial, one of the cells from the memory array was selected at a random as a target-location to be probed following the retention interval. The arrow from the memory array at the selected location was transformed according to both its orientation and colour: the orientation of the arrow was offset by either $-40^{\circ}, -15^{\circ}, -6^{\circ}, 6^{\circ}, 15^{\circ}, or\ 40^{\circ}$; likewise the colour of the arrow was offset by either $-80^{\circ}, -30^{\circ}, -15^{\circ}, 15^{\circ}, 30^{\circ}, or\ 80^{\circ}$. These increments were determined by extensive piloting of the tasks in order to try and achieve equivalent

difficulty. The transformed arrow was presented at the same location within the same 4x4 display grid until the participant gave a response or 5000ms had elapsed. A failure to respond within 5000ms was counted as incorrect. For orientation judgements, participants were provided with a reference orientation wheel to the right of the display grid, this contained directional arrows reminding the participant of the appropriate response options. Likewise, for colour judgements, participants were provided with a reference colour wheel to the right of the display grid, this contained directional arrows reminding the participant of the appropriate response options, and crucially, the colour mappings onto the circular space. Participants were required to make a two-alternative forced choice (2AFC) judgement about the direction in which the probe arrow had changed according to either its colour or orientation, relative to its counterpart arrow (same location) in the memory array. They were instructed to press the 'J' key for a clockwise response or the 'F' key for a counter clockwise response and to be both accurate and fast. In the dual response task, a probe display grid was shown first requiring one judgement type and then followed by the other (i.e. colour then orientation or orientation then colour).

Feedback

Participants were provided feedback for 300ms after each probe response by the reference wheel turning green and displaying 'correct' for a correct response, or turning red and displaying 'incorrect' for an incorrect response.

Inter trial interval

After each trial a blank 600px by 600px grid square was presented for 500ms.

Task variants

Whilst all the CDT tasks followed the same general form, what differentiated them were the required judgements about the probe arrow relative to its counterpart in the memory array grid (arrow in the same cell location). The different tasks and their judgement requirements are outlined below (see figure 4.1 for a graphical depiction).

Orientation-CDT

This task required participants to respond specifically to changes in the orientation of the probe arrow relative to its counterpart in the memory array.

Colour-CDT

This task required participants to respond specifically to changes in the colour of the probe arrow relative to its counterpart in the memory array.

Dual-CDT

This task required participants to respond to changes in both the orientation and colour of the probe arrow relative to its counterpart in the memory array. To counterbalance the order of the response judgements (i.e. colour then orientation or orientation then colour) participants were presented with two variants of this task: in one they were required to first make a judgement about the changes in colour before making a second judgement about the changes in orientation; in the other they were required to first make a judgement about the changes in orientation before making a second judgement about the changes in colour. From the participants' perspective these were presented as separate tasks, each with their own instructions and practice trials.

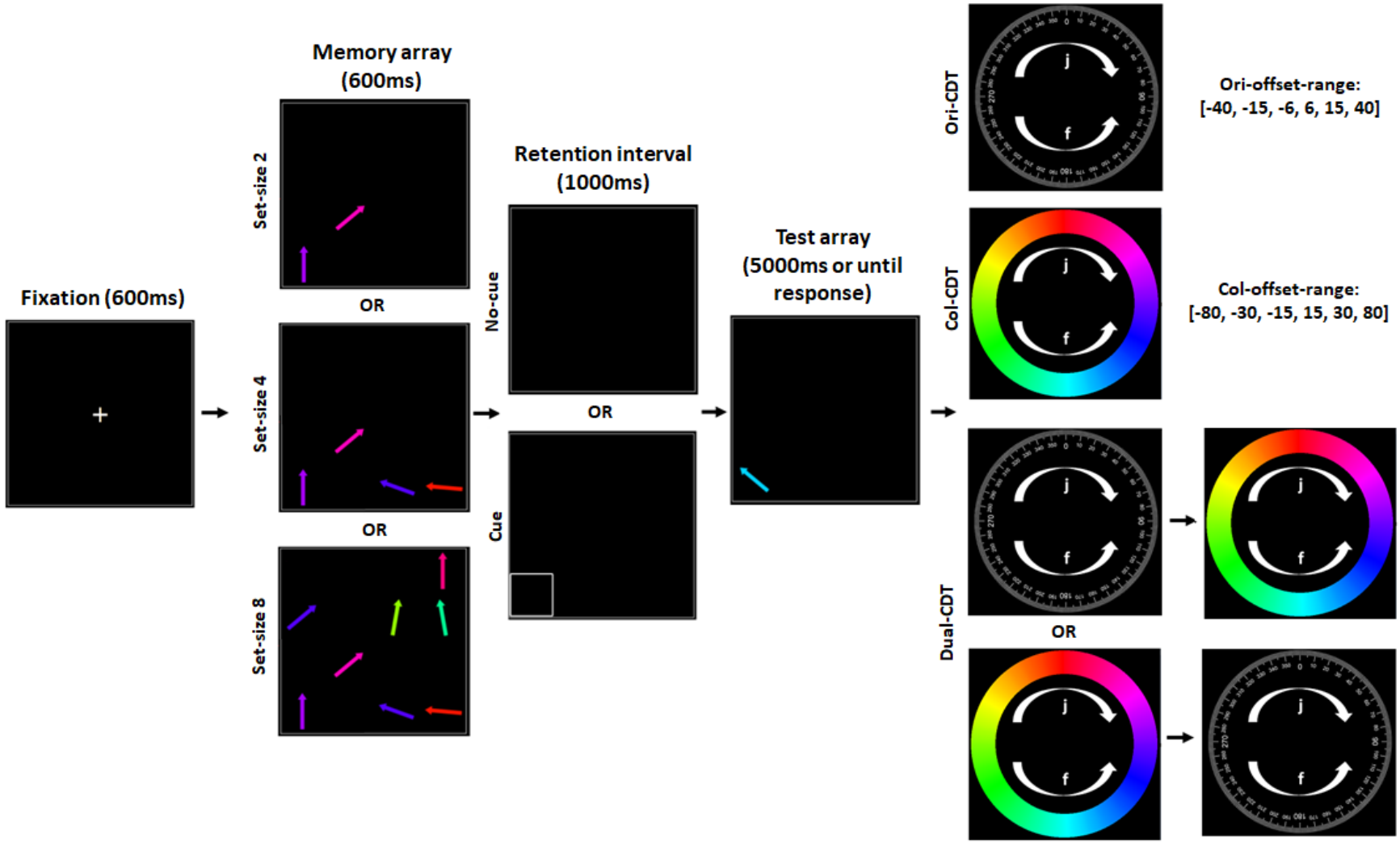
Although the stimuli parameters and trial orders were randomly generated for each CDT task independently, in the assessment tasks these were the same for all participants and at each assessment point (pre- and post- training). Trials were divided evenly according to set-size, cue-type, and offset. For both the single response tasks there were 180 trials total, split evenly into 90 retro-cue trials and 90 no-cue trials. Each cue type contained 30 trials at each set size. Each set size contained 5 trials at each offset value. For the dual response task there were 216 trials total (counterbalanced for response order and combined), split evenly into 108 cue trials and 108 no-cue trials. Each cue type contained 36 trials at each set size. Each set size contained 6 trials at each offset value.

Digit span task

This task required participants to memorize sequences of randomly sampled numbers. A number of digits ranging from 0-9 was sampled at random without replacement according to the current sequence length. For sequences >10 the initial set of 10 digits were appended with a further set of randomly sampled digits ranging from 0-9 accordingly. Participants were instructed to memorize the digit sequences as best they can and input them to a text-response box in the order they were presented. Participants received explicit instructions along with an example. They were also given three practice trials at a set size of three to help ensure they understood the task and how to respond appropriately. All participants began with a sequence length of three and were presented with up to six trials at each sequence length (maximum of

20). Participants progressed to the next sequence length if they scored over 50% accuracy at a given length, the task ended when they failed to achieve this.

Figure 4.1. Change detection task trial flowchart.



4.2.4 Training tasks and procedure

Upon completion of the first assessment session participants were randomly allocated to one of four training conditions: (1) Colour-Change-Detection-Training; (2) Orientation-Change-Detection-Training; (3) Dual-Change-Detection-Training; or (4) Digit-Span-Training. Each received specific instructions about the training phase of the study and were linked to a personalised 'homepage'. This homepage contained information about the number of sessions they had completed and how long they had to wait before starting the next, it also provided a portal to the next session when available. Participants were only allowed to access the next session after 3 hours had elapsed following completion of the previous. This homepage also contained a portal to the second assessment session, which participants were able to access 3 hours after they had completed all training sessions.

Each of the four training tasks had an almost identical assessment task counterpart (see 'Assessment tasks and procedure' for details). Although highly similar, the training tasks differed from the assessment tasks in the following ways: 1) they did not include any practice trials; 2) CDT training tasks did not contain any cue trials; 3) their difficulty was adaptive – that is, adjusted on the fly to match the performance of the participant. The difficulty level achieved by the end of one training session carried over to the next. Participants received trial-by-trial feedback, level up/down notifications, as well as feedback about the difficulty level achieved at the end of each session. The training stimuli parameters were unique to each participant; although, they followed the same parameter distributions as one another and their assessment task counterparts (i.e. same orientation/colour starting range, same orientation/colour offset range, same digit range). All participants received 12 sessions of adaptive training in total. The unique details of each training task and condition are outlined below.

Orientation-CDT-Training

This task was structurally identical to the Orientation-CDT assessment task, except the set size varied adaptively according to the participant's performance level and it did not contain any retro-cue trials. Each training session consisted of between 200-212 trials (variable due to potential level ups before the end of a block) and lasted approximately 15-17mins, with the opportunity for a short break halfway through. All participants started on the easiest difficulty level (set size of 1) in the first session, training difficulty (set-size) was then adapted using a staircase procedure, if participants scored =>75% correct over 12 trials they

were moved up a difficulty level (an increase in set size by 1) but if they scored <=58.3% correct over 12 trials they moved down a difficulty level (a decrease in set size by 1), otherwise they remained at the same difficulty level. These levels, and those for the other groups, were determined following extensive piloting of the training tasks.

Colour-CDT-Training

This task was structurally identical to the Colour-CDT assessment task, except the set size varied adaptively according to the participant's performance level and it did not contain any cue trials. Each training session consisted of between 200-212 (variable due to potential level ups before the end of a block) trials and lasted approximately 18-20mins, with the opportunity for a short break halfway through. All participants started on the easiest difficulty level (set size of 1) in the first session, training difficulty (set-size) was then adapted using a staircase procedure, if participants scored =>75% correct over 12 trials they moved up a difficulty level (an increase in set size by 1) but if they scored <=58.3% correct over 12 trials they were move down a difficulty level (a decrease in set size by 1), otherwise they remained at the same difficulty level.

Dual-CDT-Training

This task was structurally identical to the Dual CDT assessment task, except the set size varied adaptively according to the participant's performance level and it did not contain any retro-cue trials. Each training session consisted of between 200-212 trials (variable due to potential level ups before the end of a block) and lasted approximately 20-25mins, with the opportunity for a short break halfway through. All participants started on the easiest difficulty level (set size of 1) in the first session, training difficulty (set-size) was then adapted using a staircase procedure, if participants scored =>75% correct over 12 trials they moved up a difficulty level (an increase in set size by 1) but if they scored <=58.3% correct over 12 trials they were move down a difficulty level (a decrease in set size by 1), otherwise they remained at the same difficulty level. The difficulty level achieved by the end of one training session carried over to the next. The order of the response judgements (i.e. colour then orientation, or orientation then colour) alternated predictably between training sessions. On half the sessions participants were required to first make a judgement about the changes in colour of the probe arrow before making a second judgement about the change in orientation of the probe arrow; in the other half the order was reversed.

Digit-Span-Task-Training

Each training session lasted approximately 20mins (determined by a timer), with the opportunity for a short break halfway through. All participants started at the easiest difficulty level (sequence length of 1) for the first session, training difficulty (sequence length) was then adapted using a staircase procedure, if participants scored =>83.3% correct over 6 trials they moved up a difficulty level (an increase in sequence length by 1) but if they scored <=16.7% they moved down a difficulty level (a decrease in sequence length by 1), otherwise they remained at the same difficulty level. Again, these levels were determined following extensive piloting.

4.2.5 Data exclusion

All incoming data were screened for quality based on summary statistics saved using JavaScript/JATOS. Participants with particularly low accuracy and reaction times across tasks at pre-training assessment (Accuracy<53% and RT<500ms; based on pilot data) were assumed to not be engaging and excluded from the study. Furthermore, participants who did not complete all sessions were excluded from analysis. Of the 229 participants who started, 170 participants completed all sessions.

After data collection, participants who scored below 2 standard deviations (calculated task wise at pre-training) on two or more tasks at pre or post-training were excluded from all subsequent analyses. Again, this was intended to remove participants who were not engaging with the tasks. This resulted in a further 2 out of the 170 participants to be excluded, leaving 168 participants for this set of analyses.

4.3 Results

Given the complexity of the design I have included minimal descriptive statistics and focus on the primary inferential analyses of interest in this chapter. However, full summary statistics broken down by different factor combinations are provided in Appendix C.

4.3.1 Transfer effects

To investigate whether the groups show differential transfer patterns with respect to either accuracy or reaction time, I first conducted a 4 (group) x 3 (set-size) x 2 (cue-type) ANCOVA for each of the CDT task conditions (Orientation, Colour, Dual-Orientation, Dual-Colour), to test for any main effects of group, set-size, or cue-type, and their interactions, whilst co-varying for pre-training performance. To test for differences in transfer to the Digit-

Span task, I also performed a one-way ANCOVA to test for any group differences in span length, whilst co-varying for pre-training performance. The full results of these analyses, along with any other post-hoc follow ups, are provided in Appendix C. However, given that there were no group by set-size, nor group by cue-type interactions, and for purposes of brevity, in this chapter I will just report the main effects of group and associated post-hoc contrasts, at the task level (see Tables 4.1 and 4.2 and Figure 4.2). In essence, the ANCOVAs revealed that the training effects do not interact with either set size or retro-cue/no cue.

Accuracy

There was a significant main effect of group on post-training accuracy, whilst co-varying for pre-training performance, for all of the tasks: Ori-CDT ($F(3,983) = 5.703$, $p<0.001$, $\eta_p^2 = 0.017$); Col-CDT ($F(3,983) = 25.239$, $p<0.001$, $\eta_p^2 = 0.071$); Dual-Ori-CDT ($F(3,983) = 8.876$, $p<0.001$, $\eta_p^2 = 0.026$); Dual-Col-CDT ($F(3,983) = 23.983$, $p<0.001$, $\eta_p^2 = 0.068$); and Digit-Span ($F(3,983) = 31.983$, $p<0.001$, $\eta_p^2 = 0.068$). To follow up on these main effects of group, post-hoc t-tests were conducted on the adjusted (for pre-training performance) post training accuracy scores to establish precisely which group contrasts were significant (table 4.1). The significant effects for group differences are summarised below.

Orientation-CDT training vs Digit-Span training (control)

Following training, the Orientation-CDT training group had greater accuracy relative to the Digit-Span training group on the Orientation-CDT ($p=0.012$, Cohen's $d = 0.166$) and the Dual-Ori-CDT tasks ($p<0.001$, Cohen's $d = 0.271$). Conversely, the Digit-Span training group had greater accuracy relative to the Orientation-CDT group on the Colour-CDT ($p=0.045$, Cohen's $d = 0.140$) and Digit-Span tasks ($p<0.001$, Cohen's $d = 1.413$).

Colour-CDT training vs Digit-Span training (control)

Following training, the Colour-CDT training group had greater accuracy relative to the Digit-Span training group on the Col-CDT ($p<0.001$, Cohen's $d = 0.433$), Dual-Ori-CDT ($p=0.04$, Cohen's $d = 0.149$), and Dual-Col-CDT ($p<0.001$, Cohen's $d = 0.366$). Conversely, the Digit-Span training group had greater accuracy relative to the Colour-CDT training on the Digit-Span task ($p<0.001$, Cohen's $d = 1.174$).

Dual-CDT training vs Digit-Span training (control)

Following training, the Dual-CDT training group had greater accuracy relative to the Digit-Span training group on the Orientation-CDT ($p=0.033$, Cohen's $d = 0.141$), Colour-

CDT (p=0.032, Cohen's d =0.163), Dual-Ori-CDT (p<0.001, Cohen's d = 256), and Dual-Col-CDT tasks (p<0.001, Cohen's d = 0.334). Conversely, the Digit-Span training group had greater accuracy relative to the Dual-CDT training on the Digit-Span task (p<0.001, Cohen's d = 1.303).

Orientation-CDT training vs Colour-CDT training

Following training, the Orientation-CDT training group had greater accuracy relative to the Colour-CDT training group on the Orientation-CDT task (p<0.01, Cohen's d = 0.184). Conversely, the Colour-CDT group had greater accuracy relative to the Orientation-CDT training group on the Colour-CDT (p<0.001, Cohen's d = 0.576), and Dual-Col-CDT tasks (p<0.001, Cohen's d = 0.470).

Orientation-CDT training vs Dual-CDT training

Following training, the Dual-CDT training group had greater accuracy relative to the Orientation-CDT training group on the Col-CDT (p<0.001, Cohen's d = 0.301), and Dual-Col-CDT tasks (p<0.001, Cohen's d = 0.438).

Colour-CDT training vs Dual-CDT training

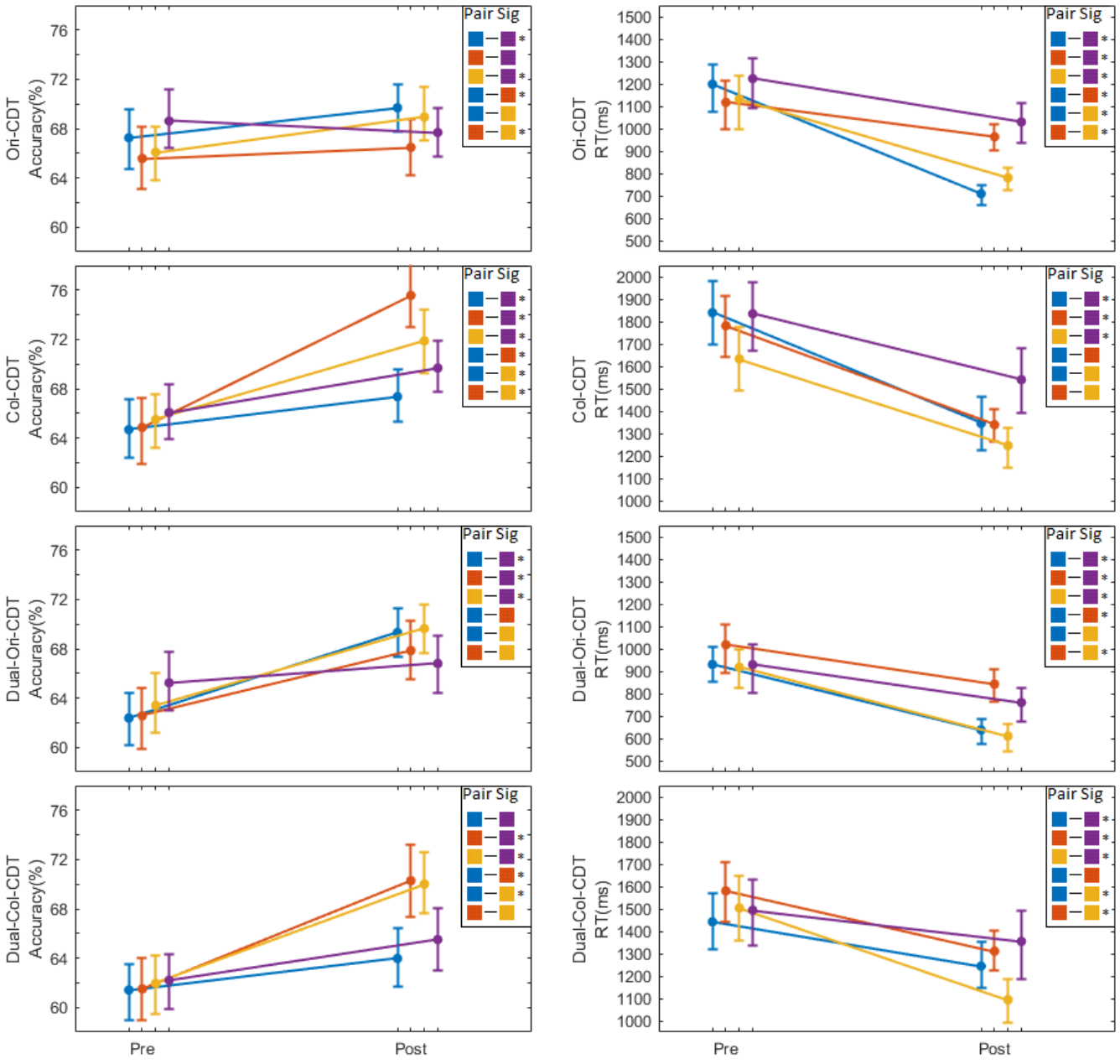
Following training, the Colour-CDT training group had greater accuracy relative to the Dual-CDT training group on the Colour-CDT task (p<0.001, Cohen's d = 0.262). Conversely, the Dual-CDT training group had greater accuracy on the Orientation-CDT task (p=0.022, Cohen's d = 0.159).

Table 4.1. Pairwise group comparisons of the whole task mean accuracy differences adjusted for baseline performance.

Task	Group contrast	Post-training accuracy (%) /span difference	t-test			
			df	*t*	*p*	*d*
Ori-CDT	Ori-Digit	2.45	84	3.024	0.0128*	0.166
	Col-Digit	-0.22	83	0.272	1.000	0.015
	Dual-Digit	2.08	83	2.547	0.033*	0.141
	Ori-Col	2.67	81	3.240	<0.01**	0.184
	Ori-Dual	0.37	81	0.447	1.000	0.025
	Col-Dual	-2.30	80	2.780	0.022*	0.159
Col-CDT	Ori-Digit	-1.95	84	2.007	0.0450*	0.140
	Col-Digit	6.22	83	6.344	<0.001***	0.433
	Dual-Digit	2.36	83	2.405	0.032*	0.163
	Ori-Col	8.17	81	8.251	<0.001***	0.576
	Ori-Dual	-4.31	81	4.351	<0.001***	0.301
	Col-Dual	3.86	80	3.874	<0.001***	0.262
Dual-Ori-CDT	Ori-Digit	3.60	84	4.491	<0.001***	0.271
	Col-Digit	2.06	83	2.559	0.04*	0.149
	Dual-Digit	3.53	83	4.379	<0.001***	0.256
	Ori-Col	1.54	81	1.892	0.176	0.123
	Ori-Dual	0.08	81	0.094	0.924	0.006
	Col-Dual	-1.46	80	1.786	0.176	0.122
Dual-Col-CDT	Ori-Digit	-1.23	84	1.351	0.354	0.095
	Col-Digit	5.07	83	5.542	<0.001***	0.366
	Dual-Digit	4.60	83	5.031	<0.001***	0.334
	Ori-Col	-6.29	81	6.809	<0.001***	0.470
	Ori-Dual	-5.83	81	6.301	<0.001***	0.438
	Col-Dual	-0.47	80	0.503	0.614	0.032
Digit-Span	Ori-Digit	-2.871	84	8.705	<0.001***	1.413
	Col-Digit	-2.335	83	7.043	<0.001***	1.174
	Dual-Digit	-2.537	83	7.613	<0.001***	1.303
	Ori-Col	0.536	81	1.605	0.330	0.329
	Ori-Dual	0.333	81	0.991	0.646	0.211
	Col-Dual	0.202	80	0.599	0.646	0.133

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

Figure 4.2. Mean accuracies and reaction times pre- and post-training for each group on each CDT task.



Note. *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected). Significant group differences are shown at post training after controlling for pre-training performance (Tables 4.1 and 4.2). Error bars show the 95% confidence interval about the mean.

Reaction time

There was a significant main effect of group on post-training RT, whilst co-varying for pre-training performance, for all of the CDT tasks: Ori-CDT ($F(3,983) = 150.37$, $p<0.001$, $\eta_p^2 = 0.314$); Col-CDT ($F(3,983) = 21.614$, $p<0.001$, $\eta_p^2 = 0.061$); Dual-Ori-CDT ($F(3,983) = 69.599$, $p<0.001$, $\eta_p^2 = 0.175$); and Dual-Col-CDT ($F(3,983) = 32.744$, $p<0.001$, $\eta_p^2 = 0.090$). The Digit-Span task was excluded from any RT analysis because there was no speeded component to it. To follow up on these main effects of group, post-hoc t-tests were conducted on the adjusted (for pre-training performance) post training RT scores to establish precisely which group contrasts were significant (table 4.2). The significant effects for group differences are summarised below.

Orientation-CDT training vs Digit-Span training (control)

Following training, the Orientation-CDT training group were faster relative to the Digit-Span training group on the Orientation-CDT ($p<0.001$, Cohen's $d = 1.270$), Colour-CDT ($p<0.001$, Cohen's $d = 0.405$), Dual-Ori-CDT ($p<0.001$, Cohen's $d = 0.513$), and Dual-Col-CDT tasks ($p<0.001$, Cohen's $d = 0.177$).

Colour-CDT training vs Digit-Span training (control)

Following training, the Colour-CDT training group were faster relative to relative to the Digit-Span training group on the Orientation-CDT ($p=0.030$, Cohen's $d = 0.140$), Colour-CDT ($p<0.001$, Cohen's $d = 0.414$), and Dual-Col-CDT tasks ($p<0.01$, Cohen's $d = 0.190$).

Dual-CDT training vs Digit-Span training (control)

Following training, the Dual-CDT training group were faster relative to the Digit-Span training group on the Orientation-CDT ($p<0.001$, Cohen's $d = 0.882$), Colour-CDT ($p<0.001$, Cohen's $d = 0.502$), Dual-Ori-CDT ($p<0.001$, Cohen's $d = 0.513$), and Dual-Col-CDT tasks ($p<0.001$, Cohen's $d = 0.579$).

Orientation-CDT training vs Colour-CDT training

Following training, the Orientation-CDT training group were faster relative to the Colour-CDT training group on the Orientation-CDT ($p<0.001$, Cohen's $d =1.471$), and Dual-Ori-CDT tasks ($p<0.001$, Cohen's $d = 0.718$).

Orientation-CDT training vs Dual-CDT training

Following training, the Orientation-CDT training group were faster relative to the Dual-CDT training group on the Orientation-CDT (p<0.001, Cohen's d = 0.533), and Dual-Col-CDT tasks (p<0.001, Cohen's d = 0.529).

Colour-CDT training vs Dual-CDT training

Following training, the Dual-CDT training group were faster relative to the Colour-CDT training group on the Orientation-CDT (p<0.001, Cohen's d = 0.948), Dual-Ori-CDT (p<0.001, Cohen's d = 0.797), and Dual-Col tasks (p<0.001, Cohen's d = 0.528).

Table 4.2. Pairwise group comparisons of the whole task mean reaction time differences adjusted for baseline performance.

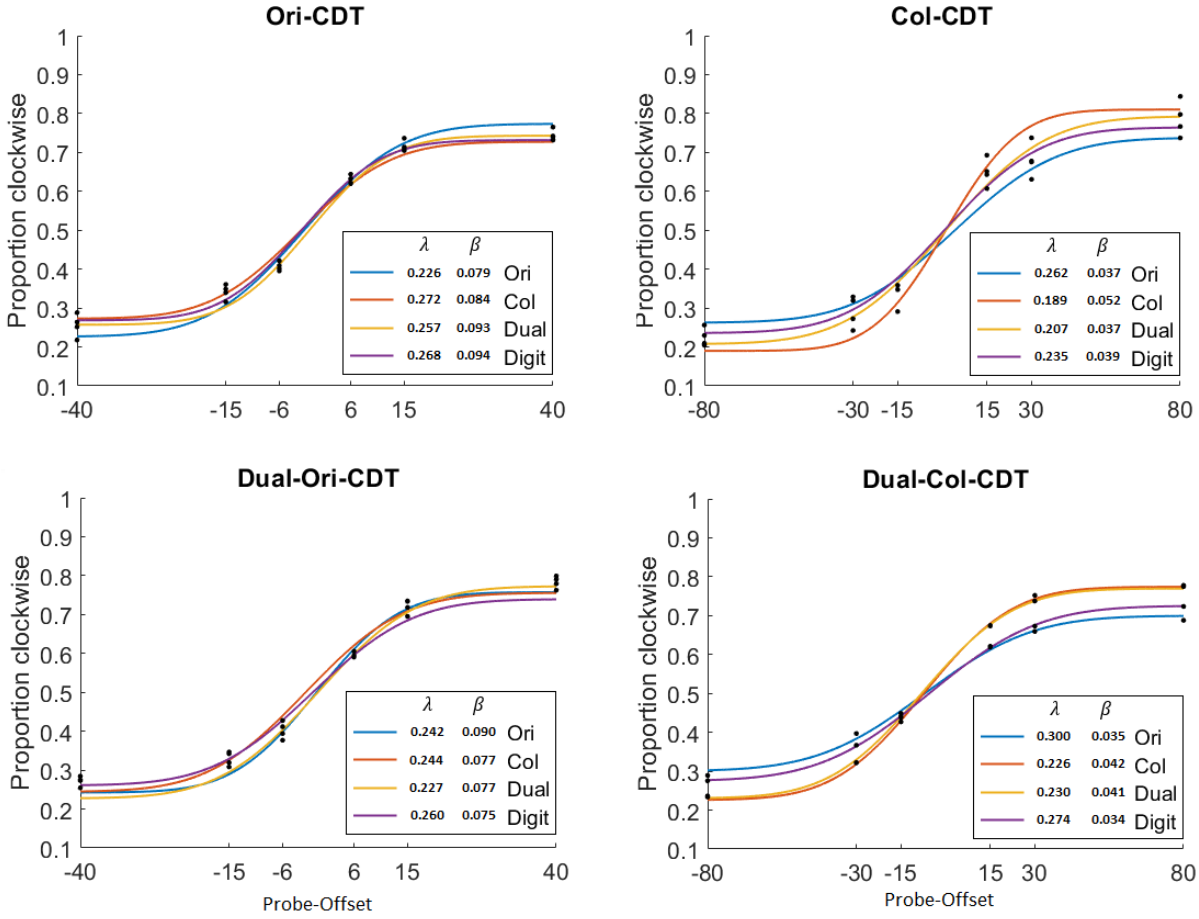
Task	Group contrast	Post-training reaction time difference (ms)	Between group t-test			
			df	*t*	*p*	*d*
Ori-CDT	Ori-Digit	-313.97	84	18.285	<0.001***	1.270
	Col-Digit	-37.54	83	2.164	0.030*	0.140
	Dual-Digit	-224.64	83	12.962	<0.001***	0.882
	Ori-Col	-276.42	81	15.781	<0.001***	1.471
	Ori-Dual	-89.33	81	5.104	<0.001***	0.533
	Col-Dual	187.09	80	10.644	<0.001***	0.948
Col-CDT	Ori-Digit	-197.28	84	6.321	<0.001***	0.405
	Col-Digit	-184.19	83	5.861	<0.001***	0.414
	Dual-Digit	-225.84	83	7.123	<0.001***	0.502
	Ori-Col	-13.09	81	0.411	0.747	0.036
	Ori-Dual	28.55	81	0.890	0.747	0.078
	Col-Dual	41.65	80	1.296	0.584	0.138
Dual-Ori-CDT	Ori-Digit	-123.68	84	8.184	<0.001***	0.513
	Col-Digit	38.97	83	2.551	0.021*	0.144
	Dual-Digit	-146.05	83	9.604	<0.001***	0.513
	Ori-Col	-162.65	81	10.530	<0.001***	0.718
	Ori-Dual	22.37	81	1.454	0.146	0.113
	Col-Dual	185.03	80	11.894	<0.001***	0.797
Dual-Col-CDT	Ori-Digit	-83.82	84	3.059	<0.01**	0.177
	Col-Digit	-88.68	83	3.212	<0.01**	0.190
	Dual-Digit	-266.45	83	9.671	<0.001***	0.579
	Ori-Col	-4.86	81	0.173	0.862	0.013
	Ori-Dual	182.63	81	6.547	<0.001***	0.529
	Col-Dual	177.77	80	6.332	<0.001***	0.528

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

4.3.2 Psychometric functions

In addition to comparing groups on mean accuracy and RT, I also wanted to investigate whether group differences in training outcomes on the CDT tasks were driven by an increase in the number of, or the quality of, item representations held in memory. To do so, I fit a cumulative Gaussian curve models to the distributions of clockwise responses over the range of target-stimulus offsets for each group, on each task (collapsed across set-size and cue-type), according to the methods outlined in Murray et al (2013), using the Palamedes toolbox (Prins & Kingdom, 2018). Each of these models provided two key parameters of interest: λ (Lambda) – the asymptote of the curve, providing an estimate of the probability that a probe item was represented in memory; and β (Beta) – the slope of the curve, providing an estimate of the precision of a probe item representation. The models and the corresponding parameters of interest are shown in figure 4.3 below.

Figure 4.3. Cumulative Gaussian functions fit to each of the CDT tasks for each group across trials.



The functions were fit to the group as a whole post-training, combining across trials from all participants. This was to try and maximize the number of trials and improve the overall fit of the model (Although see discussion; Murray et al., 2013, Prins & Kingdom,

2018). The statistical comparisons between groups were then conducted by using 500 bootstraps to create confidence intervals for each group. These could then be used to compare the groups statistically (whilst correcting for multiple comparisons; Table 4.3). The Colour CDT group produced significantly more precise representations, and showed a greater asymptote, relative to the Orientation CDT group, when performing the Colour CDT assessment. The Orientation CDT group showed a higher asymptote relative to either the Colour CDT group or the controls, when performing the Orientation CDT. Interestingly, the Orientation CDT group also showed a poorer asymptote relative to Dual CDT and Colour CDT trainees when performing the Colour CDT task (either alone or in its Dual CDT manifestation). One possibility is that the Orientation CDT training actually interfered with the Colour CDT task.

Table 4.3. Pairwise group comparisons of the whole task differences on key psychometric parameters

Task	Group contrast	Differences			
		B	*p*	*λ*	*p*
Ori-CDT	Ori-Digit	-0.016	0.504	-0.042	0.036*
	Col-Digit	-0.011	0.808	0.004	0.592
	Dual-Digit	-0.002	0.704	-0.013	0.654
	Ori-Col	-0.005	0.704	-0.046	0.04*
	Ori-Dual	-0.013	0.600	-0.029	0.360
	Col-Dual	-0.008	0.808	0.017	0.654
Col-CDT	Ori-Digit	-0.002	1.000	0.027	0.324
	Col-Digit	0.013	0.136	-0.045	0.096
	Dual-Digit	-0.002	1.000	-0.030	0.324
	Ori-Col	-0.015	0.036*	0.072	<0.001***
	Ori-Dual	-0.001	0.744	0.056	0.060
	Col-Dual	0.014	0.060	-0.015	0.288
Dual-Ori-CDT	Ori-Digit	0.015	0.250	-0.018	0.890
	Col-Digit	0.002	1.000	-0.017	0.888
	Dual-Digit	0.001	1.000	-0.034	0.504
	Ori-Col	0.014	0.264	-0.001	0.492
	Ori-Dual	0.014	0.144	0.016	0.890
	Col-Dual	0.001	0.896	0.017	0.672
Dual-Col-CDT	Ori-Digit	0.001	0.796	0.026	0.324
	Col-Digit	0.008	0.384	-0.049	0.104
	Dual-Digit	0.007	0.352	-0.045	0.198
	Ori-Col	-0.007	0.384	0.074	0.012*
	Ori-Dual	-0.006	0.352	0.071	0.020*
	Col-Dual	0.001	0.796	-0.003	0.486

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

4.3.3 Correlations pre- and post- training

To further explore the relationships between tasks and how these might change as a function of training, I examined the correlations at pre assessment, post assessment, and the

difference between the two (Table 4.4). This mirrors the analysis conducted previously in Chapters 2 and 3. To establish whether correlations changed significantly following training I used a permutation method wherein I randomly sampled (n=2000) the pre and post assessment task performances for each group and calculated the pairwise changes in the correlation coefficients to estimate a distribution and produce p-values. There was a significant difference on the Col-CDT versus Dual-Col-CDT for the Dual-CDT training group.

Table 4.4. Pairwise comparisons of the changes in correlations between change detection tasks following training within each group

Task-pair	Group	Correlations			
		Pre	Post	Difference	*p*
Ori vs Col	Ori-CDT	0.474	0.555	0.081	0.261
	Col-CDT	0.614	0.523	-0.091	0.285
	Dual-CDT	0.514	0.636	0.123	0.158
	Digit-Span	0.541	0.463	-0.077	0.260
Ori vs Dual-Ori	Ori-CDT	0.744	0.809	0.065	0.247
	Col-CDT	0.814	0.818	0.004	0.470
	Dual-CDT	0.676	0.680	0.004	0.480
	Digit-Span	0.685	0.806	0.121	0.075
Col vs Dual-Ori	Ori-CDT	0.320	0.465	0.145	0.170
	Col-CDT	0.746	0.526	-0.220	0.056
	Dual-CDT	0.445	0.442	-0.003	0.454
	Digit-Span	0.624	0.582	-0.041	0.352
Ori vs Dual-Col	Ori-CDT	0.490	0.605	0.114	0.224
	Col-CDT	0.594	0.683	0.090	0.230
	Dual-CDT	0.535	0.644	0.108	0.206
	Digit-Span	0.474	0.572	0.097	0.254
Col vs Dual-Col	Ori-CDT	0.715	0.713	-0.002	0.487
	Col-CDT	0.755	0.806	0.051	0.267
	Dual-CDT	0.521	0.857	0.336	<0.001***
	Digit-Span	0.737	0.624	-0.112	0.121
Dual-Ori vs Dual-Col	Ori-CDT	0.570	0.655	0.084	0.271
	Col-CDT	0.801	0.694	-0.107	0.078
	Dual-CDT	0.592	0.549	-0.043	0.321
	Digit-Span	0.643	0.753	0.111	0.181

Note. *p < .05. **p < .01. ***p < .001

One of the shortcomings the previous chapter was a lack of between group comparisons in the task-pair correlation changes. Here, I also compared the differences between groups (Table 4.5). To establish whether the between group contrasts for correlation differences following training were significant, I used the permuted pre and post samples from the above analyses to form a single distribution of the differences and estimated p-values according to the proportion of the distribution above or below 0 (two-tailed). There was a significant difference between the Ori-CDT and Col-CDT training groups on the Col-

CDT versus Dual-Ori CDT task pair, with the two task becoming relatively less correlated for the Col-CDT group. There were also a significant differences between Dual-CDT and all the other training groups on the Col-CDT versus Dual-Col-CDT task pair, with the correlation between these two tasks becoming relatively increased in each case.

Table 4.5. Group contrasts for the pairwise changes in correlations between the change detection tasks following training

Task-pair	Group contrast	Difference in r-change	*p*
Ori vs Col	Ori-Digit	0.158	0.201
	Col-Digit	-0.013	0.464
	Dual-Digit	0.200	0.121
	Ori-Col	0.172	0.199
	Ori-Dual	-0.042	0.384
	Col-Dual	-0.214	0.146
Ori vs Dual-Ori	Ori-Digit	-0.055	0.293
	Col-Digit	-0.116	0.157
	Dual-Digit	-0.116	0.201
	Ori-Col	0.061	0.336
	Ori-Dual	0.061	0.333
	Col-Dual	-0.001	0.463
Col vs Dual-Ori	Ori-Digit	0.186	0.147
	Col-Digit	-0.179	0.184
	Dual-Digit	0.039	0.429
	Ori-Col	0.365	0.046*
	Ori-Dual	0.147	0.224
	Col-Dual	-0.218	0.181
Ori vs Dual-Col	Ori-Digit	0.017	0.499
	Col-Digit	-0.008	0.481
	Dual-Digit	0.011	0.480
	Ori-Col	0.025	0.486
	Ori-Dual	0.006	0.490
	Col-Dual	-0.019	0.475
Col vs Dual-Col	Ori-Digit	0.110	0.197
	Col-Digit	0.163	0.086
	Dual-Digit	0.449	<0.001***
	Ori-Col	-0.053	0.350
	Ori-Dual	-0.338	0.011*
	Col-Dual	-0.285	0.017*
Dual-Ori vs Dual-Col	Ori-Digit	-0.027	0.439
	Col-Digit	-0.218	0.054
	Dual-Digit	-0.154	0.163
	Ori-Col	0.191	0.119
	Ori-Dual	0.127	0.220
	Col-Dual	-0.064	0.380

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

4.4 Discussion

Motivated by increasing evidence for transfer specificity, in the previous chapter I presented a high-powered online training study that examined transfer patterns within a set of hierarchically nested perceptual discrimination tasks. Transfer effects were bound to a specific judgement type (Spikiness). In a similar vein, this chapter used a high-powered online study to examine transfer patterns following training. This time I used a set of nested CDTs. This chapter and the previous are heavily overlapping in both motivation and approach but there are some key differences: Firstly, whilst both sets of tasks involve aspects of perceptual discrimination, memory, and other executive functions, the CDT tasks would generally be considered 'higher level' due to all the tasks involving more memory items and more specific response requirements, both of which impose greater resource demands. Secondly, whilst both sets of tasks are hierarchically nested, containing tasks that include all features of their lower-level constituents, here I opted to use fewer tasks with less variation between them (only judgement type compared to both judgement and presentation) but greater variation within (i.e. set-size, cue-type, offset). In turn, this also allowed me to have the complete set of training conditions (all assessment tasks had a training counterpart). Overall, the design used in the previous chapter allowed me to ask general questions about feature relationships and transfer. The design used in the current chapter allowed me to ask more specific, theoretically motivated questions about transfer, its boundary conditions, and potential underlying mechanisms. Specifically, I asked the following questions: 1) Does training lead to the acquisition of skills that enhance the number of items stored in memory, the precision of those items, or both? 2) Does training interact with the spatial allocation of attention? 3) Are the skills acquired during training in the single judgement tasks bound to their specific judgement types of colour and orientation, or do they transfer to one another? 4) Do the skills acquired during training in the single task conditions transfer 'up' the task hierarchy to a dual judgement (Orientation and Colour) task and vice versa?

The broad pattern of data showed both accuracy and RT improvements following training, with each assessment task showing an effect of group. That is, for each task, the degree of improvement shown, relative to pre-training baseline, was moderated by the kind of training participants had undertaken. However, there were no significant interactions with set-size or cue type (no cue or retro-cue trials), so for all following analyses I collapsed across these factors. There were a number of interesting group-specific training effects. For example, each CDT training group made significantly greater on-task gains, relative to the

active control group. Participants who underwent colour CDT or orientation CDT training gained significantly on their respective assessment tasks, but there was no significant transfer between them. On the contrary, the orientation training group were *significantly worse* than the control group at the colour CDT assessment. Rather than recapitulate each result here, because there are so many, I will group the key results around the four research questions I attempted to answer with this study.

4.4.1 Does training lead to the acquisition of skills that enhance the number of items stored in memory, the precision of those items, or both?

Previous research has shown attentional and/or experiential modulation of both the quantity (Buschkuchl et al., 2017; Griffin & Nobre, 2003; Norris et al., 2020; Moriya, 2019; Heur & Schubo, 2016) and quality (Heur & Schubo, 2016; Bays et al., 2009; Marshell & Bays, 2013; Moriya, 2019) of VWM representations. As such, I predicted that training would enhance on task-performance with respect to both the number of item representations held in memory and the precision of those item representations. I tested this in two ways: 1) varying the number of items (set-size) in the memory array and 2) varying the degree of offset for both the colour and the orientation of the probe stimulus relative to its counterpart in the memory array.

If training primarily enhanced capacity, then we would expect the CDT training groups to outperform the control group disproportionality on the higher set-sizes compared to the lower set-sizes. However, despite training being aimed at increasing capacity, there were no significant group by set size interactions on any of the tasks, suggesting that by this measure at least, capacity has not been increased. Varying the probe offsets allowed me to fit psychometric functions to the data and derive two key parameters of interest, for each group on each of the tasks. The first indicates sensitivity to changes in offset (Beta, i.e. quality of the representation), and the other indicating the threshold for the number of items held (Lambda). First, with respect to the Beta parameter, none of the CDT-training groups showed greater sensitivity to change relative to controls either on or across tasks (although the effects for both the Col-CDT and Dual-Col-CDT conditions were trending in this direction, they did not remain significant after correcting for multiple comparisons). Second, with respect to the Lambda parameter, the Ori-CDT training group showed a significantly higher capacity threshold relative to controls on-task, suggesting that training on the Ori-CDT task does in fact engender skills related to capacity. However, neither of the Col-CDT nor Dual-CDT training groups showed significantly higher capacity thresholds, on- or across tasks, relative

to controls (although again, these effects were trending in that direction but failed to survive multiple comparisons).

Further to the comparison with controls, there were some specific effects for these parameters when comparing between the training groups. For example, on the Col-CDT task, the Col-CDT training group showed both enhanced precision and an enhanced capacity threshold relative to the Ori-CDT group. Relatedly, both the Col-CDT and Dual-CDT training groups showed an enhanced capacity threshold relative to the Ori-CDT group. Conversely, the Ori-CDT training group showed a greater capacity threshold on the Ori-CDT task compared to the Col-CDT. These findings mirror those from the whole task accuracy comparisons and suggest that there are some specific training effects associated with the judgement type trained. Moreover, given that these effects only come out between training groups and not relative to control, indicates that not only are some of the skills acquired task specific but also that there may be some skills acquired that are detrimental outside of the original training context. In other words, training on the Orientation CDT may make your Colour CDT slightly worse, at least relative to the control group.

Taken together, the above findings suggest that there are some task-specific effects pertaining to both quantity and quality of item representations. However, these effects are patchy at best, and reconciling them becomes tricky when integrating them with the far more robust basic accuracy effects between groups. One explanation for why we might find performance differences between the CDT training groups and controls on simple accuracy measures, but not in the precision or capacity estimates from the psychometric modelling, is that training affects both the quality and quantity of item representations. When these are essentially aggregated with crude overall accuracy measures, they are more robustly detected, whereas when they are segregated into modelling components the effects wash out when we control for multiple comparisons. Relatedly, due to time constraints I was only able to fit the psychometric functions in a coarse manner across set sizes at post-training. This means that any effects are likely dulled by virtue of including data from the supra-threshold set size eight, which are incredibly noisy and do not produce reliable model convergence when taken by themselves. This is further amplified by not having access to the full data due to the server error. With more data and/or focusing solely on data from set-sizes two and four, the same analysis may allow us to disentangle effects on precision versus capacity more convincingly. Moreover, comparing across groups post training does not account for any potential pre-

training differences, as in the other analyses. This will also need to be accounted for in the full analysis.

To summarise this first question: Based upon the available information at the time of writing, there is no compelling evidence that would allow us to conclude that CDT training significantly impacts capacity or precision selectively. There are no interactions with set-size, and relative to controls only one significant improvement to the asymptote of the psychometric functions. There are no significant improvements in the slope of the psychometric functions, relative to controls.

4.4.2 Does training interact with the spatial allocation of attention?

Top-down attentional modulation has been shown to impact both the encoding and maintenance of CDT item representations (Posner, 1980; Griffin & Nobre, 2003; Astle et al., 2012). Prior to data collection, I tentatively predicted that training would disproportionately affect processes associated with the encoding/early maintenance (i.e. those prior to cue onset) phase of the trial due to the relatively long presentations times used here, which may leave more room for strategically modifying encoding efficiency (Gaspar et al., 2013; Vogel at el., 2006), and evidence suggesting an increased value of placing cues early on at encoding/maintenance (Griffin & Nobre, 2003; Astle et al., 2012). An alternative possibility is that training boosts the allocation of spatial attention during maintenance.

To distinguish these two possibilities, I included a retro-cue on half of the assessment CDT task trials, halfway (500ms) through the maintenance phase of the trial to indicate the position of the upcoming probe stimulus. The idea being that if training disproportionately affected processes earlier on in the trials, then we would expect any training effects to persist regardless of the cue, relative to the control group. In other words, cue type should not interact with training group. Alternatively, if the training affected later processes, then the training gains might be negated by the cue – on cue trials everyone can orient their top-down attention regardless of what training they have had. In other words, we would get an interaction between group and cue type – the training gains are only present on no-cue trials.

In line with previous findings, there was a main effect of cue-type for both accuracy and RT, with cue-trials both enhancing accuracy and decreasing RT (Souza & Oberauer, 2016; Griffin & Nobre, 2003; Astle et al., 2012), on all tasks except for accuracy on the Ori-CDT task where the effect was null. However, crucially, there were no group by cue-type interactions, indicating that whatever was learnt during training was unaffected by the

presence of the cue. One possible interpretation is simply that whatever is gained during training has nothing to do with the spatial orienting of attention during the maintenance phase. If it did then we would expect it to interact with the cueing effect. This may suggest that whatever is enhanced by the training reflects some other mechanism, either early at encoding, or late during retrieval, but it does not pertain to the allocation of attention during VWM maintenance.

4.4.3 Are the skills acquired during training in the single judgement tasks bound to their specific judgement types of colour and orientation, or do they transfer to one another?

Prior to data collection, I predicted that skill gains associated with *enhancing the quality* of memory representations would be judgement-specific, but that any gains stemming from *enhanced capacity* would be transferable. The first half of this prediction was based upon increased saliency for the trained features. There is evidence suggesting that memory recall for within-object features can fail independently in continuous response paradigms (Bays et al., 2011; Fougnie & Alvarez , 2011). The second half of this prediction was based upon evidence suggesting that maintaining a greater number of items requires within-object features to be bound as integrated items (Alvarez & Cavanagh, 2004; Bays et al., 2009; Bays & Hussain, 2008; Oberauer & Hein, 2012; Luck & Vogel, 2013; Luria & Vogel 2011). Moreover, a prior study had shown transfer between orientation and colour judgement types in a CDT paradigm following colour CDT training in terms of overall capacity.

However, as previously discussed, I had difficulty parsing quantity vs quality training effects. Despite a tentative indication that some effects may be more driven by skills pertaining to quantity (i.e. asymptote), I now consider this question across both of these at the whole task level with respect to both accuracy and RT. In the accuracy data (i.e. when comparing mean accuracy across groups) the data speak strongly to the gains being feature-specific. Being trained to remember an increased number of items in terms of their orientation does not improve your memory for their colour, and vice versa. Thus, we did not replicate the findings of Norris et al. 2020. On the contrary, there was instead evidence for 'negative transfer' – training on the orientation variant makes you worse at the colour variant, at least relative to the controls. There was a hint of this in the psychometric function data also. One possibility is that this is driven by some inhibitory process; training on the orientation variant may actively encourage participants to supress the interfering colour information that is bound within the stimulus, or that the orientation information is biased to such an extent that it becomes interfering when the colour information is relevant. A key

difference between this study and that of Norris et al. (2020) is that I used feature bound items, such that the perceptual characteristics of the memoranda were matched across all CDT tasks. In contrast, Norris et al. used items that contained colour information or orientation information, but never both. Another difference is that I varied the offset of the probe stimulus, whereas Norris et al. left theirs constant and at an easily discriminable level. As mentioned above, one possibility is that varying the degree of discriminability in this study encouraged a very feature-specific strategy which may have prevented transfer. These factors may explain the differences between the two studies.

The RT data show far more widespread improvements. Training on any variant of the CDT makes you faster at all other variants of the CDT task. There are also some additional improvements more specific to the Orientation CDT and Dual CDT group. Both groups are faster than the Col-CDT group when orientation judgements are required in either a single judgement or dual judgement context. Similarly, the Dual-CDT group is quicker on colour judgements than other groups for colour judgements in the dual condition. These effects are hard to interpret because improved speed at a task could result from any number of processes. However, one likely possibility is that CDT training enhances retrieval speeds and/or motor responses. The response required in all the CDT variants is to use a colour wheel or orientation wheel to decide whether the stimulus had rotated clockwise or anti-clockwise. This is quite a bizarre thing for participants to do, and a strong possibility is that this response process itself is trainable. Because all CDT variants use this response method, getting faster will likely transfer across tasks. This may be why the RT data appear to show such generic transfer effects.

4.4.4 Do the skills acquired during training in the single task conditions transfer 'up' the task hierarchy to a dual judgement (Orientation and Colour) task and vice versa?

Whilst I did not make any specific predictions about the directionality of transfer in at pre-registration, given the findings in the previous chapter it seems reasonable now to assume that we would observe transfer 'down' the hierarchy, but not transfer 'up'. That is, we might observe improvements in the simpler CDT variants from having trained on the dual version, but not vice versa. The accuracy data demonstrate that in some cases transfer along the hierarchy can be bidirectional. Those who trained on the dual variant of the CDT also gained significantly on the simpler feature-specific versions of the tasks, relative to controls. Those who trained on the colour variant of the CDT also gained significantly on the dual version (with a similar effect size to the dual trainees), and whilst this was most prominent for the

colour feature of the dual variant, the gains were also significant for the orientation version, relative to controls. In other words, there may not be significant transfer between orientation and colour variants, but colour CDT trainees do get better at the orientation elements of the dual CDT. Meanwhile, the orientation CDT trainees do not show this generalisation. They only improve on the orientation elements of the dual CDT (again with a similar effect size to the dual trainees). It thus seems from the accuracy data that training on either colour or orientation CDT boosts performance wherever you encounter that stimulus type, and to roughly the same extent as those who have trained on the dual version (in terms of effect size). Likewise, training on the dual version makes you better on either of the simpler versions, again with a similar effect size to those who train selectively on those respective variants. Interestingly however, colour CDT trainees show transfer to orientation only within the context of the dual CDT. One possible explanation is that the gains for the colour variant are simply much bigger. If the colour variant becomes much easier for those who have encountered it during training, then it may free up resources for those participants to allocate to the orientation feature of the dual task. In essence, the training gains themselves are somewhat asymmetric, and this may explain the apparent asymmetry of transfer.

4.4.5 Correlational relationships following training

In both previous chapters there were several changes in pairwise task correlations within-group following training. By contrast, here only one task pair changed within a single group: following Dual-CDT training there was a substantial increase in association between the colour CDT task and the dual colour CDT, perhaps indicating that they now recruit more common processes. However, it is curious that we do not see the same gain in association for Colour-CDT training group. One possibility is that the additional executive demands of the dual condition interfere with the colour CDT groups' usual on-task processes – in essence, because these participants have not encountered the dual condition, other than in the assessments, it still involves additional processes and thus the correlation between the colour CDT variants (alone versus in the dual context) is smaller.

One shortcoming in the previous chapter was a lack of between-group comparisons for these correlation effects. However, here I also compared these changes between groups. In accordance with the above effect, the Dual-CDT training group see a greater strengthening of association between the Col-CDT task and the Dual-Col-CDT task, relative to the other groups. Further to this, when contrasted against one another we see a weakening of association between the Col-CDT and Dual-Ori-CDT tasks for the Col-CDT training group

relative to the Ori-CDT training group for whom we see a slight strengthening of association. This is perhaps further evidence for feature specific training gains, which drive strong correlations between variants that share the feature, and for some mild interference caused by judgement types other than those trained in the single judgment conditions.

4.4.6. Limitations

There are several limitations to the current study that are important to consider when interpreting the findings. Firstly, I was missing some of the assessment data due to a technical error with the JATOS server. This will be rectified when the issue is resolved. In practice, this means that the current analyses are slightly underpowered relative to the original design (I am currently missing 10% of the participants' assessment data). Secondly, the psychometric modelling includes all trial types, even though we know that the set size 8 trials are incredibly noisy. A more optimal design would have been to avoid these trials altogether, but I had worried we would get ceiling effects at set size 4 following training. Nonetheless, it is likely a good idea to remove them from the psychometric function calculation. Thirdly, ideally, I would have more trials such that I could fit good psychometric functions for individual participants. In the current analysis I fit the functions at a group level, with bootstraps used for group-wise comparisons. When the full dataset is available, I will attempt to fit these functions to individual subjects, excluding the set size 8 trials, both before and after training for a proper comparison of modelling components.

4.4.7 Conclusions

By training VWM using three variants of a CDT task, alongside an active control group, I was able to answer a number of key questions about patterns of transfer. Firstly, training on either simple variant of the CDT does not transfer to the other. On the contrary, training on the orientation version may make you somewhat worse at the colour version. Second, transfer patterns are bidirectional within the nested hierarchy. Training on the dual version of the CDT makes participants almost as good at the simple versions, as if they had trained selectively on just those tasks. Likewise, training on the simple versions makes participants better at those elements of the more complex task. In the case of colour CDT trainees, they show improvements on the orientation elements of the dual CDT task, suggesting between-variant transfer only in the context of the dual task. Third, whatever is being trained in CDT it is unlikely to be the spatial allocation of attention during the maintenance period. There are no interactions between training type and cue type. Whatever

is enhanced by the training it likely happens either pre-cue, such as during encoding, or during retrieval. Finally, I was not able to conclude as to whether training enhances capacity or precision, because the results from the psychometric function analysis do not clearly support one or the other.

Chapter 5: General Discussion

5.1 Background and purpose of this thesis

Variation in the outcomes and conclusions between cognitive training studies – leading to uncertainty and controversy - is in large part driven by fundamental differences in research questions, methodologies, availability of resources, and chance. Early findings in the field have since been heavily scrutinised and attributed to methodological shortcomings, such as: a lack of statistical power, failing to correct for multiple comparisons, or a lack of appropriate control groups (Sala & Gobet, 2019; Simons et al., 2016). It appears that much confusion has also been created by comparing apples to oranges and contrasting apples with apples, so to speak; or put more formally, due to the two intimately related and fundamental issues of task impurity – the extent to which any given task measures an intended construct – and task similarity – the extent to which two tasks overlap with one another (Taatgen, 2013; Gathercole et al., 2019; Smid et al., 2020). This essentially can mean that in some cases we misinterpret changes following training.

Task design is the vehicle by which researchers make inferences about cognitive processes, which are otherwise unobservable. As such, operationalising task relationships is perhaps the most fundamental and important issue in the cognitive sciences (Maul, 2016). In fact, it could be argued that operationalising task relationships is the goal of the cognitive sciences. Nowhere is this more pertinent than in the field of cognitive training, a field shrouded by controversy, at the very heart of which lies the deceptively simple idea that something learnt in the context of one task may transfer to that of another. Defining how two tasks relate to one another is a prerequisite for making quantitative predictions about transfer and answering any ‘what?’ or ‘how?’ questions of its nature (Reder & Klatzky, 1994; Barnett & Ceci, 2002; Taatgen, 2013; Gathercole et al., 2019; Smid et al., 2020).

As has been heavily emphasised throughout, how one decides to quantify and talk about relationships between tasks strongly influences the interpretation of training outcomes and any potential applications (Reder & Klatzky, 1994; Barnett & Ceci, 2002; Simons et al., 2016; Taatgen, 2013; Gathercole et al., 2019; Smid et al., 2020). For example, a popular approach early in the field (like that of the studies contained in Chapter 2), was to deliver a varied training diet, with multiple tasks thought to measure general abilities such as working memory (WM), or related executive functions, and then test people across multiple other tasks also purporting to measure other general abilities, without strictly quantifying the

relationships within or between them. In many cases, what is touted as domain general enhancements are likely just examples of near transfer, because the assessment task is also in the training battery. Whilst broad training approaches may have their own merits, this 'blunderbuss' approach can lead to what would now be considered an exaggerated view of the benefits of training, about such training regimes enhancing general abilities (Redick, 2019; Sala & Gobet, 2019). This is in stark contrast to more recent evidence suggesting that not only do typical training regimes fail to enhance general abilities, but often fail even to produce transfer between two tasks that are almost identical and differ subtly by only a single, seemingly arbitrary feature (Simons et al., 2016; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019).

Despite its importance, there remains no generally accepted method for defining how tasks relate to one another and consequently no taxonomy by which to determine the extent to which two given tasks overlap. In short, how do we know how 'near' or 'far' they are from one another (Taatgen, 2013)? Without an established taxonomy of the tasks ('Taskonomy'), it is difficult to pin down precisely which processes are targeted by, or manifest because of, training and any boundary conditions for their transfer (Simons et al., 2016; Gathercole et al., 2019). Producing a cognitive taxonomy is by no means trivial and requires a level of theoretical convergence that is currently lacking (but see Taatgen, 2013). In the absence of a cognitive taxonomy, shortcomings of correlational approaches in mind, and increasing amounts of evidence that typical transfer effects are tied to specific features; task analytical, feature based, approaches provide a promising alternative for taxonomizing task relationships and exploring training effects (Gathercole et al., 2019).

In particular, there is a call for more systematic and tightly controlled manipulations of task features to better establish potential boundary conditions of transfer and advance cognitive theory (Katz et al., 2018; Redick, 2019; Sala & Gobet, 2019; Von Bastian & Oberauer, 2014; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Taatgen, 2013). In response, and in an attempt to circumvent some of the problems arising from the issues of task impurity and task similarity, much of the approach taken and language used in this thesis, represents a conceptual shift away from organising tasks by broad hypothetical constructs and towards organising them instead by the specific task features from which they are comprised. This was the primary motivation behind the experimental Chapters 3 and 4. However, as I have also emphasised throughout, other approaches also have strong merit, are not mutually exclusive, and ought to be used in tandem and integrated where possible.

A related issue is how task relationships are conceptualised and realised *across individuals* (Karbach et al., 2017). This was in part the motivation for the work presented in Chapter 2 of the thesis. The other goal of that chapter was to explore the use of a novel methodological technique. New multivariate approaches to thinking about task relationships and profiles of change following training may also be needed to advance the current understanding of the field.

Broadly, this thesis set out to explore how different conceptualisations of task relationships inform training induced transfer effects and their interpretation. The discussion that follows outlines some of the ways in which the thesis has addressed the following questions I posed at the end of Chapter 1:

i. Are unsupervised machine learning algorithms a viable multivariate alternative for representing, analysing, and interpreting cognitive training data?
ii. Does training alter task relationships?
iii. Are there subgroups of participants with different profiles of change following training?
iv. What types of task relationships best predict transfer patterns following training on different tasks within nested feature-based hierarchies?
v. Are transfer patterns following training on different tasks within nested feature-based hierarchies directional with respect to feature complexity?

5.2 Task relationships and transfer

In an attempt to remedy the often hard-to-define task relationships and hard-to-identify training effects found in broader training regimes containing several different complex tasks, Chapters 3 and 4 of the thesis presented two large, randomised, and tightly controlled online training studies. In both cases, the tasks were systematically varied and hierarchically organised with respect to their extrinsic feature combinations, identified a-priori. This approach allowed me to unambiguously quantify task similarity with respect to task features and use specific features as potential constraints to transfer. Arranging the sets of tasks hierarchically also allowed me to ask questions about complexity and the direction of transfer cascades.

Specifically, the experiment presented in Chapter 3 explored transfer effects following training on two tasks within a set of six nested perceptual discrimination tasks. Whilst all tasks involved making same-different judgments between two spikey shapes, the

task features varied with respect to judgement type (number of spikes or ‘spikiness’), presentation type (simultaneous or delayed), and task switching. One training task was a relatively low-level paradigm that involved making mono-judgement decisions about the spikiness of two shapes presented simultaneously. The other training task involved switching between the two judgement types for two shapes presented with a short delay in-between. For the Simultaneous-Spikiness training group both judgement type and switching context constrained transfer but not presentation type. For the Delayed-Switching training group judgement type also constrained transfer and presentation type was a constraint for transfer but only in a switching context. Across both groups, the best predictor of overall transfer was whether an assessment task shared a single specific feature (spikiness). The overall proportion of shared features was predictive of transfer following training on a task lower down in the hierarchy but not following training on a higher-level task despite this producing more widespread transfer. Finally, between tasks correlations were not predictive of transfer for either group and subject to change following training.

Chapter 4 was similar in its conception, but this time I attempted to focus even further, identifying the potential mechanisms and boundary conditions of transfer effects within a set of three nested change-detection-tasks (CDTs). Not only did I systematically vary features between tasks (Judgement types: orientation, colour, and both), I also systematically varied them within task (set-size, cue-type, and change offset), to try and unearth some more specific mechanisms that might affect training outcomes. Each assessment task had a complimentary training condition, resulting transfer effects for each were compared against an active control group and to one another. There was no positive transfer between either of mono-judgement tasks. In fact, those who trained on the Ori-CDT performed worse than controls on the Col-CDT. Both groups demonstrated vertical transfer ‘up’ the hierarchy for their respective on-task judgement types in a dual judgment context. Interestingly, the Col-CDT training group showed bilateral transfer upwards. That is, they showed orientation judgement improvements relative to controls but only in a dual-judgement context. Transfer patterns were bidirectional within the task hierarchy, those who trained on the dual judgement CDT task performed almost as well as those who trained solely on those tasks (although there were still some task specific benefits for training on the Col-CDT). With respect to reaction time, there were widespread transfer effects across tasks for all CDT training groups relative to control as well as some task specific effects relative to one another. Given that there were no group-by-cue-type interaction effects it seems unlikely that CDT training affected the

allocation of attention during maintenance. Finally, the extent to which training effected the quality or quantity of item representations was unclear.

Taken together these findings echo those from other recent studies and provide further demonstration that even within sets of closely related tasks there appear to be constraints on transfer associated with specific features (Soveri et al., 2017; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019). They further demonstrate that these constraints are not well accounted for by any simple conceptualisations of task relationships: constraints are variable with respect to conditions, sets of tasks, and task complexity. For example, whilst the transfer patterns in Chapter 3 were best predicted by the presence of a single feature, this only accounted for a small amount of the variance and differed between groups. Likewise, in Chapter 4, transfer to the mono-judgement tasks was constrained by judgement type for both groups but not in context of a dual-judgement task for one of the groups. Moreover, in contrast to the same-different training in Chapter 3, the CDT training in Chapter 4 transferred both up and down, to and from, a more complex and executively demanding dual-judgement context. This highlights the need for more nuanced and detailed interpretations of transfer – not only is transfer tied to specific features but also specific contexts. Potential feature-based constraints such as judgement type or switching requirements present in one set of tasks are not the same as those in another set of tasks.

Relatedly, these findings also stress the fact that transfer effects are always to be considered relative to a control group and whatever impact their training had for cognitive processes recruited in any given assessment. For example, relative the Ori-CDT group those who trained on Digit-Span may have experienced less interference on the Col-CDT because the ‘task-set’ or ‘cognitive-routine’ employed by the Ori-CDT group was specialised to perform better on orientation judgements but was nonetheless triggered by the externally identical features of the Col-CDT task (Rogers & Monsell, 1995; Dreisbach & Wenke, 2011; Gathercole et al., 2019). On the other hand, the same set of process that were a hindrance in the context of a mono-judgement task may be of help relative to the Digit-Span group in a dual-judgement context, as participants were more familiar with the general task flow, they had more attentional resources available compared with the Digit-Span group to deal with the additional executive demands of switching. This reasoning extends not only to comparisons with the ‘control’ group but also to those between the other training groups, who essentially act as controls in their own right, and help explain task the task specific effects, that is the differences in magnitude or direction of effect sizes.

5.3 Individual differences and transfer

In chapter 2 of the thesis, I presented the use of two unsupervised machine learning algorithms (SOMs and K-means-clustering) as alternatives for modelling multivariate training data toward exploring individual differences and training outcomes more generally. This involved combining data across several studies on four popular measures of WM, some of which contained pre- and post-training data for children who had undergone a broad training regime aimed at WM and related cognitive abilities. SOMs proved capable of effectively representing multivariate data across the four tasks. The patterns of change observed in these SOM representations implied that processes drawn upon to perform the tasks may have changed following training. Furthermore, K-means clustering identified four distinct performance profiles to which those children that had undergone training were assigned before and after training. Changes in group membership revealed differentiable improvement trajectories, predictive of performance on a separate measure of fluid intelligence. The approach taken here provides a viable alternative for representing, analysing, and interpreting cognitive training data that goes beyond changes in individual tasks and allows for the exploration of individual differences in training trajectories.

The presence of both different cognitive profiles prior to training and different trajectories of change in profile following training highlights the need for further investigations into individual differences that go beyond univariate approaches and other a-priori demographic groupings, toward multivariate approaches that group participants based on performance. This resonates with the message coming from previous investigations look at treatment by aptitude effects (Karbach et al., 2017; Guye et al., 2017; Smid et al., 2020). In the same way that different training interventions produce different outcomes and task relationships, so too do the lives of individuals, each a unique experience akin to a lifelong training programme, naturally leading to differences in training outcomes.

Potential age related and motivational factors may also worth considering in the current thesis. In the experimental design of the work of chapters 3 and 4, I purposefully recruited young adults in the age range of 18-30 as these are seen to be relatively stable population and I wanted to constrain potential developmental differences. Moreover, in both studies I control for differences in baseline performance somewhat by including it as a covariate. In chapter 2 however, I combined samples of children across age ranges and abilities. The fact that tasks used in the online training studies were gamified somewhat, contained ongoing feedback, and provided monetary incentives may have effected

participant's motivation and the way in which they interacted with both the training and assessment tasks. Given that both the factors of age and motivation have previously been shown to play a role in training outcomes (Karbach et al., 2017; Green & Bavelier., 2008), I perhaps haven't given these factors as much attention as they deserve, and they may well have contributed to findings within and between the studies presented here.

5.4 Limitations

There were several limitations with the approaches taken here that are important to consider when interpreting the findings here and of training outcomes more generally. As previously mentioned, chapter 2 combined across many different samples, ages, and abilities. This will have inevitably introduced more variance that may contribute to some of the findings. Although, in some respects this may also be an advantage when looking at individuals. Further, in the training sample, participants were trained across a broad range of tasks. Whilst some have argued that broad training regimes may bring about more generalisable – longer lasting benefits (Green & Bavelier, 2008; Klingberg, 2010), they also make is difficult to identify any precise mechanisms responsible for training effects because they could be due to any number of processes affected by the training, as has already been discussed at length. Interpretability of effects was further compromised by the fact that the training and assessment tasks were not explicitly and systematically mapped onto one another, making hard to disentangle task specific effects from those that may be more general (Holmes et al., 2019; Norris et al., 2019; Smid et al., 2020).

The primary limitation of the study presented in Chapter 3 was that in did not contain the full range of training groups making it difficult to say exactly what exactly constrained transfer in some situations. For example, including a group that trained exclusively on the enumeration judgement would help determine whether a lack of transfer to this judgement type was due to a lack of experience or task sensitivity. Likewise, including a group that trained exclusively on the Delayed-Spikiness task would help determined the extent to which transfer to and from this measure was limited by presentation type.

Across both training studies, participants only had a limited amount of time to train (especially in Chapter 3), as we opted to allocated resources to sample size instead. It may be that some training effects take much longer periods of time to manifest and vice versa. I suspect this may be one reason why the cognitive training literature can seem so counter-intuitive to our day-to-day notions of transferable skills. Intuitively, we believe that different

domains of expertise give rise to skills that transfer more readily. For example, we would reasonably expect that a long-distance runner would be better at sprinting than a chess player, and that a chess player would be better at draughts than a long-distance runner. Of course, we cannot rule out potential modulating factors because of a lack of experimental control. Nonetheless, perhaps it takes thousands of hours rather than just a few to generate generalizable transfer effects. The longest training study to date was the Cogito study conducted by Schmiedek et al. (2010), who used 100 sessions of training, and the data from which was analysed and briefly covered in Appendix A. However, there were large ceiling effects present in the data that made this hard to interpret. Also, for some types of generalisations the opposite may be true. That is, novelty may be required for something to generalise, as is emphasised by Gathercole et al. (2019).

Finally, Chapter 4 was missing some of the data, meaning that training gain analyses were unavailable, and the current analyses were not at full power. Second, the psychometric functions were only fit in a coarse manner across participants, cue-types, and set sizes within group at post-training. Ideally these would be fit in a more specific manner following the removal the some of the noisier data found at a set size of 8. Also, whilst having such complex within task designs by varying set-size, offset, and cue type, potentially allows for the identification of more specific processes responsible, it may also muddy the water and reduce statistical power. Moreover, varying the offset degrees in the training tasks may have biased what was learnt during training and again makes it more difficult to interpret transfer effects.

5.5 Future directions

The main sentiment expressed here, is similar to that of others (Katz et al., 2018; Redick, 2019; Sala & Gobet, 2019; Von Bastian & Oberauer, 2014; Gathercole et al., 2019; Holmes et al., 2019; Norris et al., 2019; Simons et al., 2016; Taatgen, 2013; Smid et al., 2020). That is, future cognitive training research needs to more theory driven, systematic, nuanced, and higher resolution in its approach. Primarily, there is an overarching need for the field to come together and work collaboratively toward formalising and agreeing upon task relationships in the form of hierarchical taxonomies. I believe a good starting point would be to do this first with respect to task features across differing levels of granularity. However, the possible permutations are infinite, so researchers would have to restrict this space considerably. Moreover, these also need systematically mapping onto process-based accounts that factor in individual differences. Further, models should consider that task relationships

are volatile and subject to change as a function of training. At a minimum, future training studies should provide explicit, quantifiable, and systematic accounts for how both the training and assessment tasks recruited relate to one another. They should also make theory driven predictions about the expected outcomes and aim to power the studies accordingly. As such, computational modelling approaches appear to be an extremely promising avenue of research for cognitive training. These force researchers to make both the task and learning parameters explicit. Moreover, computational models can be trained on varying amounts of tasks for varying amounts of time without incurring additional costs other than computing power. Researchers can than examine how task relationships within the models change as a function of training. Finally, training studies typically use a pre-train-post set up. However, many training/transfer effects are likely happening on a much smaller scaler. One idea would be to use high powered short burst training studies – i.e., one round of exposure to one task, the next on another, and so on. This would enable researchers to look at transfer at a much higher temporal resolution. Perhaps the goal of typical training programmes should not be to enhance any higher order cognitive processes but instead to understand lower order processes in the early stages of task learning.

5.6 Concluding remarks

Cognitive tasks do not measure the same thing in the same person at two different points in time nor the same thing in two different people at the same point in time.

Appendices

Appendix A – Supplementary methods and analyses to Chapter 2

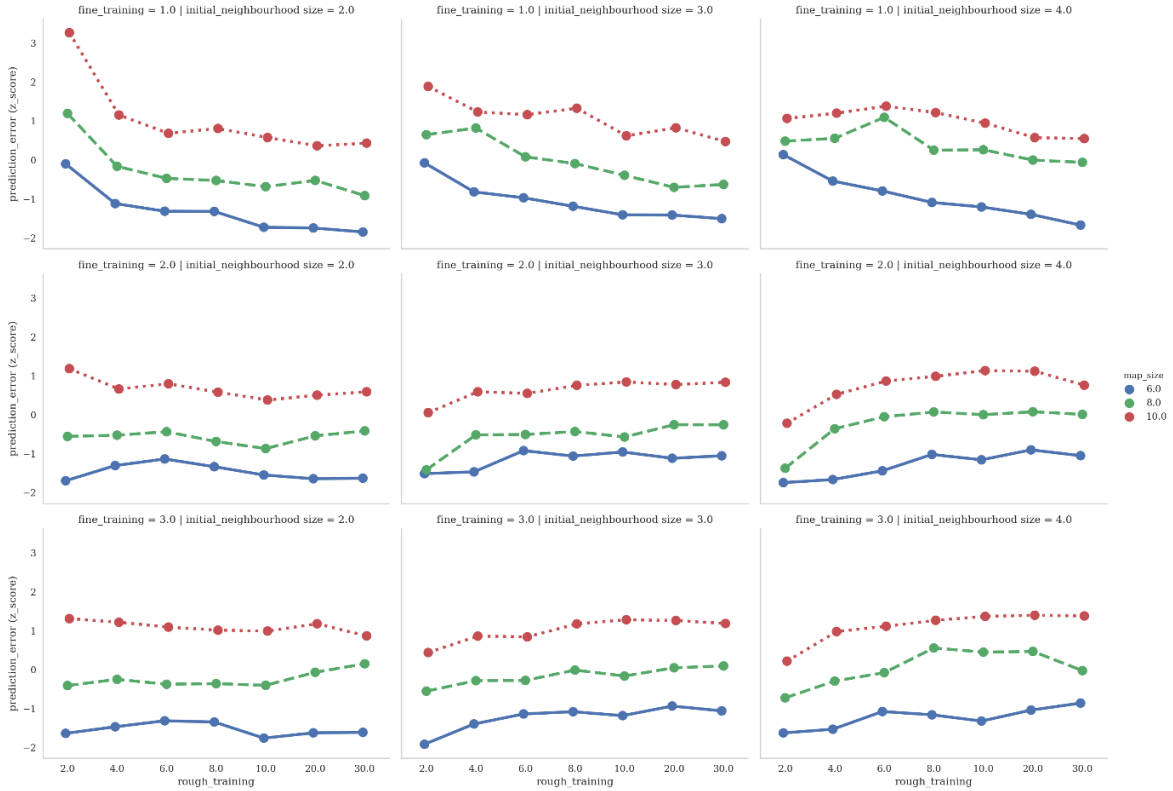
Selection of SOM parameters

I tested a range of SOM parameters provided by the Matlab 2017a Neural Network Toolbox. These included the map size, initial neighbourhood size, ordering phase steps and fine-tuning phase steps. I used a composite of quantization error and prediction error to evaluate each combination of the parameters within the tested range. Quantization error is defined as the mean absolute distance between the input vectors (i.e. training data) and their corresponding Best Matching Unit (BMU), which is an indicator of how well the model represents the input data. As discussed in the Method section, prediction error, defined as the mean absolute distance between the predicted and true values from the reserved testing data, is an indicator of the model's ability to generalise to unseen data. Hence, I combined these two measures towards the aim of representing the input data whilst maintaining generalisability. Specifically, quantization errors and the mean prediction errors across the 4 cognitive measures were standardized with respect to their own distribution, achieved from all possible combinations of model parameters within the testing range, before being summed. For each combination of parameters, the results averaged over 100 iterations were used. The range of each parameter tested over are as follows: Map size: 6, 8, 10; Initial neighbourhood size: 2, 3, 4; Ordering (rough training) phase steps: 2, 4, 6, 8, 10, 20, 30; Fine-tuning phase steps: 1, 2, 3.

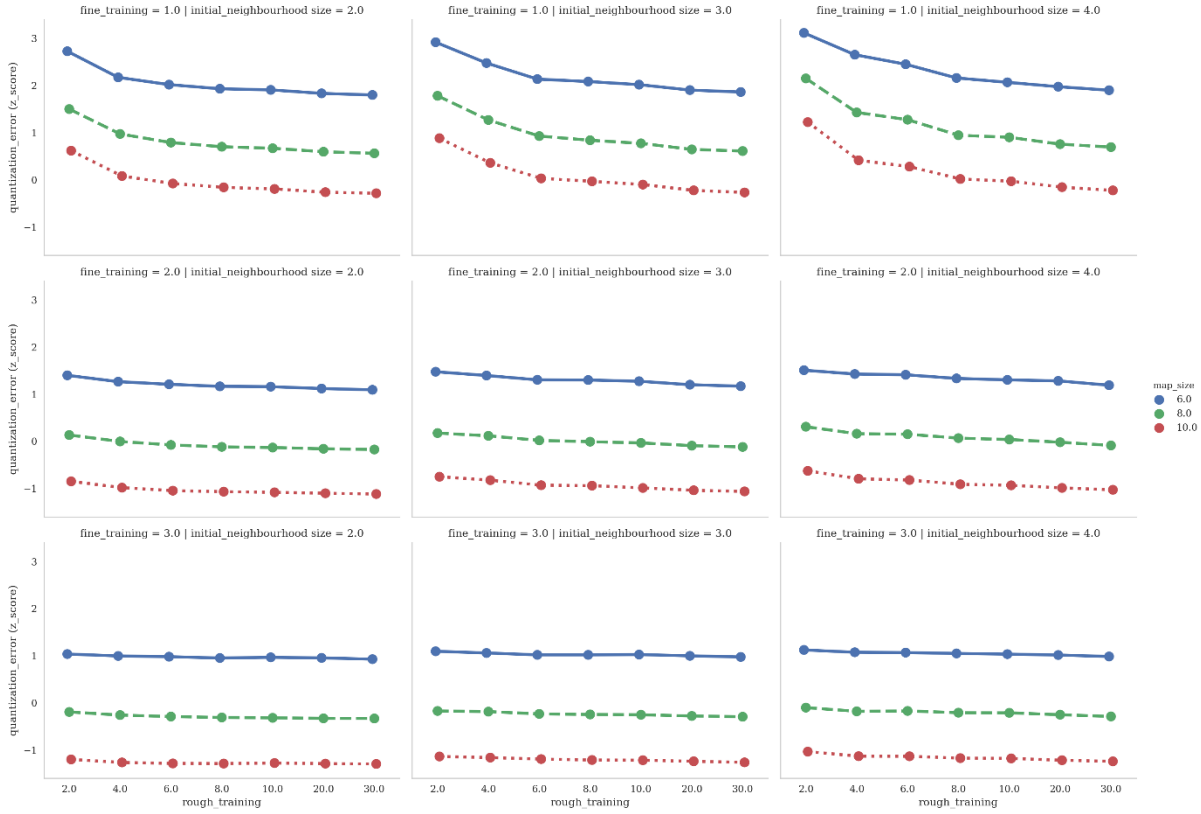
Figures A.1 and A.2 capture some of the consequences of different parameter combination selections with regards to prediction error, quantization error, and composite score respectively. The effect of SOM map size (i.e. number of nodes) exhibited the most influence on the former two measures (Figure A.1a and A.1b) which were in opposite directions: a larger map led to lower quantisation error but higher prediction error. On the other hand, the effect of the ordering phase was only distinguishable below the range of 10 iterations. As such, it was not surprising that the composite score demonstrates a more ambiguous picture. I chose the combination of map size =10, ordering phase =10, fine-tuning phase = 2, initial neighbourhood = 2 as it produced the least composite error in the range of parameters we initially tested. Subsequently, I extended the range tested over to include ordering phase steps of 2,4,6,8 and 15. This revealed 2 parameter combinations that were marginally better but most likely negligible with regards to how they affect the findings.

Figure A.1. Overview of (a) mean prediction error and (b) quantisation error

(a)

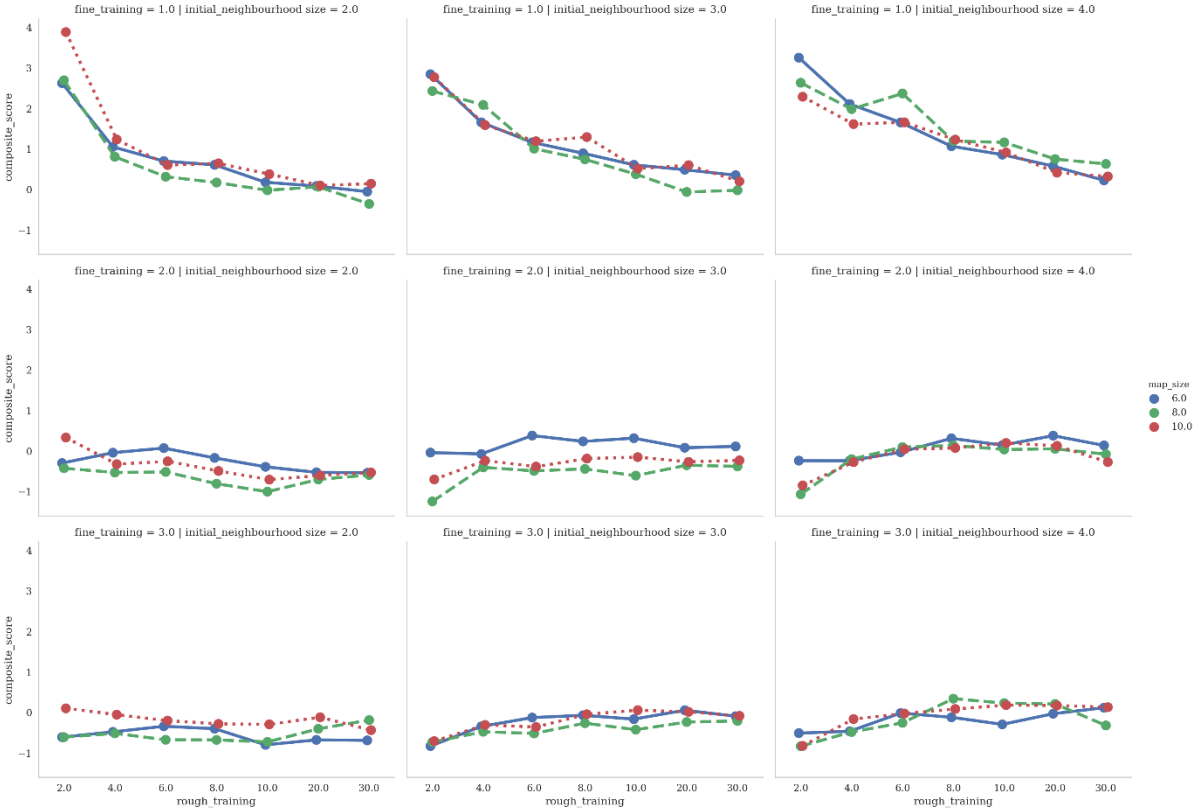


(b)



Note. Lower scores represent better outcomes.

Figure A.2. Composite score as function of SOM training parameters.



Note. Lower scores represent better outcomes.

Selection of K in identifying subgroups with differential profiles

The choice of K is somewhat arbitrary but could potentially influence the cognitive profiles associated with each subgroup. I chose K=4 as it resulted in a statistically meaningful grouping of participants, as well as producing relatively homogenous improvement trajectories in the cognitive training data. Below are several figures that show the resulting profiles for different numbers of K and their resulting silhouette values (a way of determining robustness of grouping). I also show the resulting profiles from using a K of 4 upon the Pre and Post training data, the results of which show that despite some variation they are somewhat similar. Further, I include a comparison between the K-means on the SOM-weights vs those on the raw data. The silhouette coefficient is a measure of how close each data point is to its own cluster compared to the neighbouring clusters and thus provides a way to assess parameters like the number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients near +1 indicate that a point is far away from the neighbouring clusters and therefor properly clustered. A negative value indicates that those samples are not robustly clustered. As Figure A.3 shows, applying k-means clustering on the raw dataset results in a lower average silhouette coefficient compared to SOM weights, and some individuals were incorrectly assigned to certain subgroups, as indicated by negative silhouette scores. This highlights the fact that the SOM reduces the noise in the data, thereby making the clustering more robust.

As can be seen in the Figure A.3, a two-cluster solution was favourable in terms of robustness. However, I opted for a four-cluster solution in the end as it also appeared a reasonably stable solution, whilst also allowing me capture more information and nuance with regards to the task performance profiles in the data (see figure A.4).

Figure A.3. Silhouette values for each cluster (K = 4) and the averaged silhouette coefficient (orange line) on SOM weights and raw CALM/ACE data respectively.

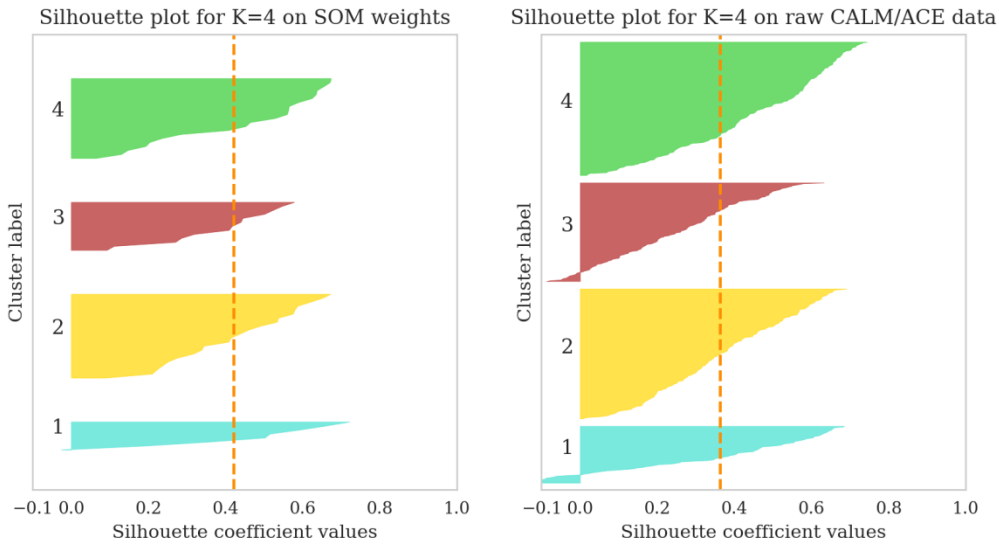
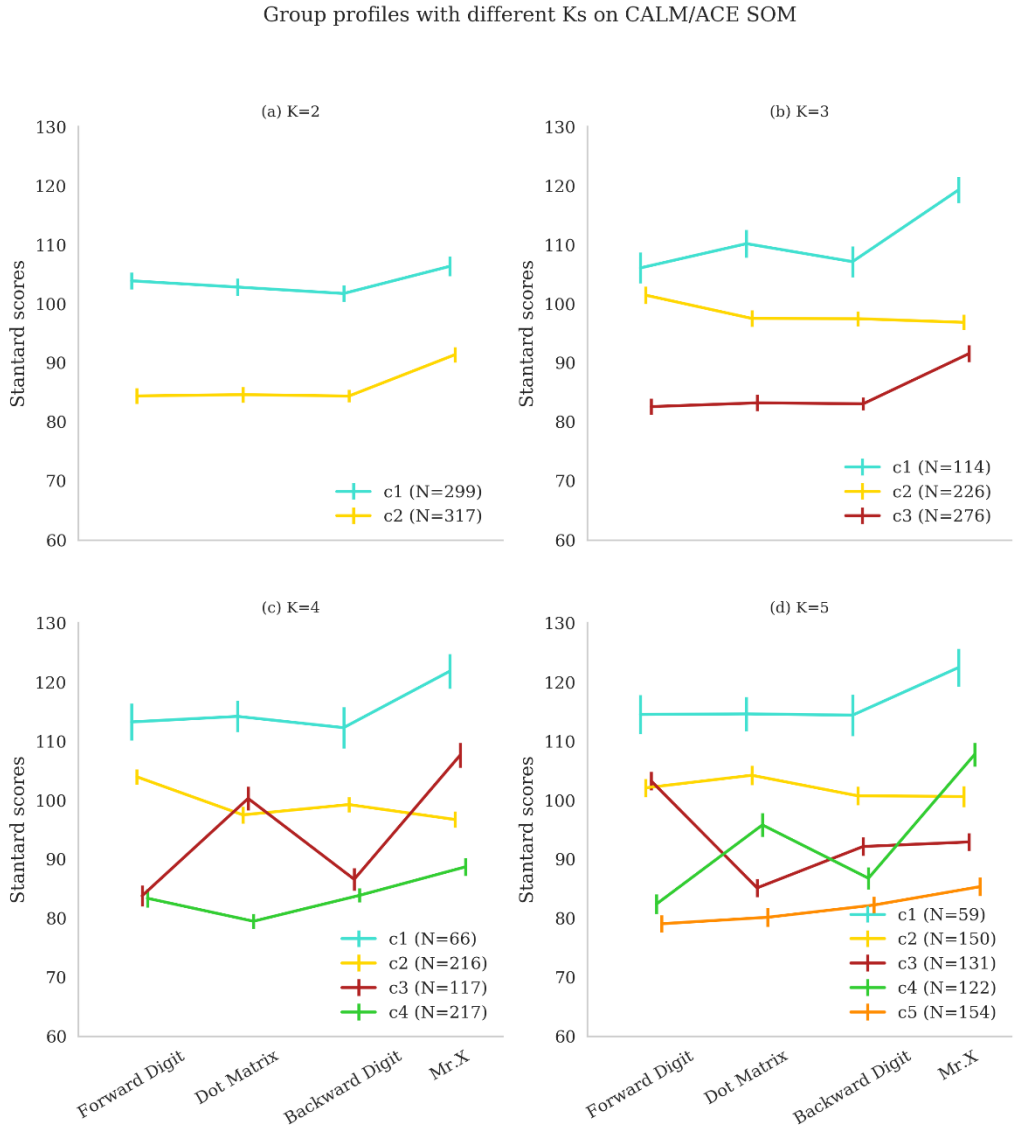


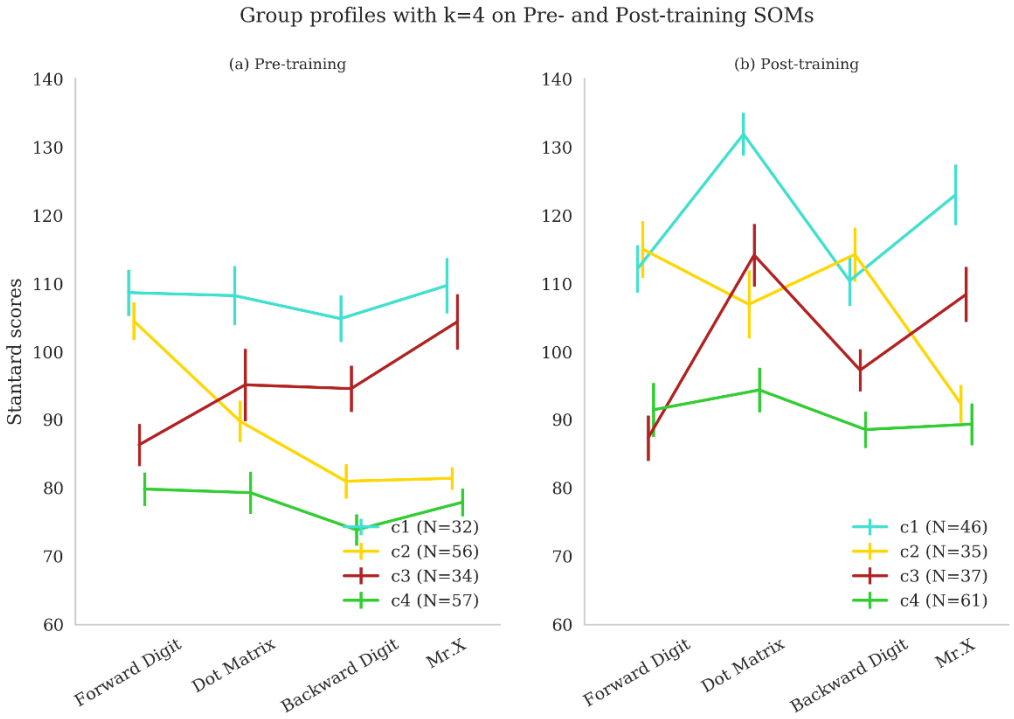
Figure A.4. Task performance profiles in the CALM-ACE dataset for differing number of K.



Although I didn't do an in depth analysis, you can see that the profiles derived from a clustering solution applied directly to the SOM models trained on the Pre-Training data (Figure A.5a) are quite distinct from those derived from the CALM/ACE data (see main text Figure 2.4d). This suggests that the two samples are somewhat different in nature, but this difference may in part be an artefact of biased sample size, since the ACE dataset consisted of 90 typically developing children whose cognitive profiles were likely to be absent in the pre-training population as a result of the inclusion criteria. In fact, it can be seen that the scores of the highest-performing group in Figure A.5a were on average lower, but the group size is bigger than those of the highest-performing group in Figure 2.4d. In other words, the pre-training sample lacks the high performance profiles that existed in the CALM/ACE and post-training samples. On the contrary, profiles derived from a clustering solution applied to

the SOM map trained on the Post-Training data (Figure A.5b), are similar versions of those derived from the CALM-ACE data clusters (see Figure 2.4e).

Figure A.5. Task performance profiles for a K of 4 on SOMs fit to the Pre- and Post-Training datasets respectively.

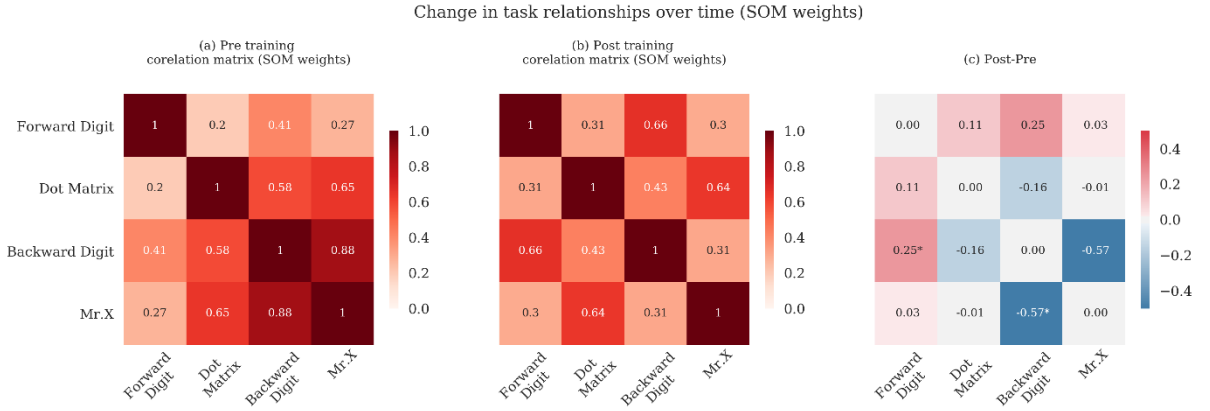


SOMs can detect changes in task relationships not present in the raw data

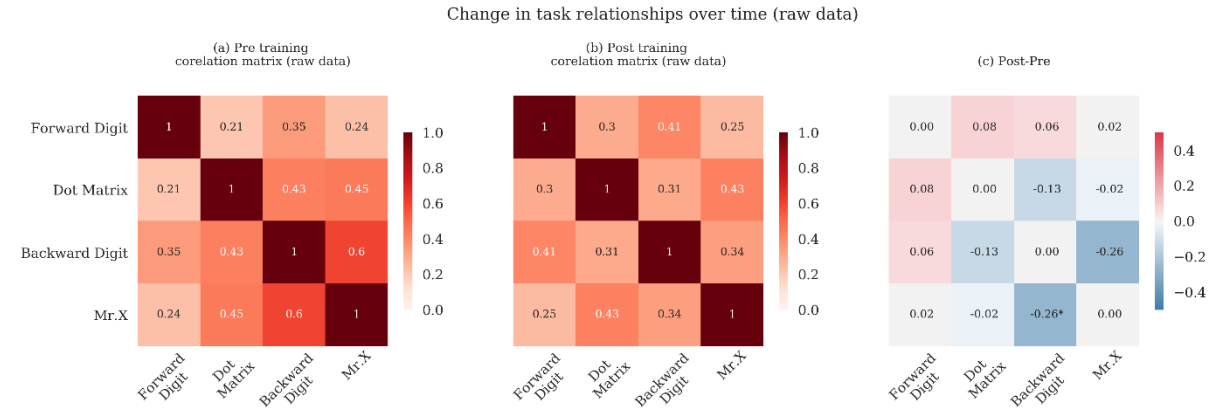
This section highlights the contribution of SOM as a representative and noise-tolerant model of the original data. Specifically, the noise-reducing property of the learning algorithm lends itself to strengthening the relationships and the changes thereof that are present in the raw data (Yin, 2008). Using the model, two pairwise relationship changes were identified (Figure A.6a and also see Figure 2.3 in main text), whereas only one (Backward digit-Mr. X) was found by computing raw cross-correlations, but with attenuated magnitude (Figure A.6b).

Figure A.6. Between task correlations at Pre-Training, Post-Training, and the difference between the two; a comparison between SOM weights (a) and raw data (b).

(a)



(b)



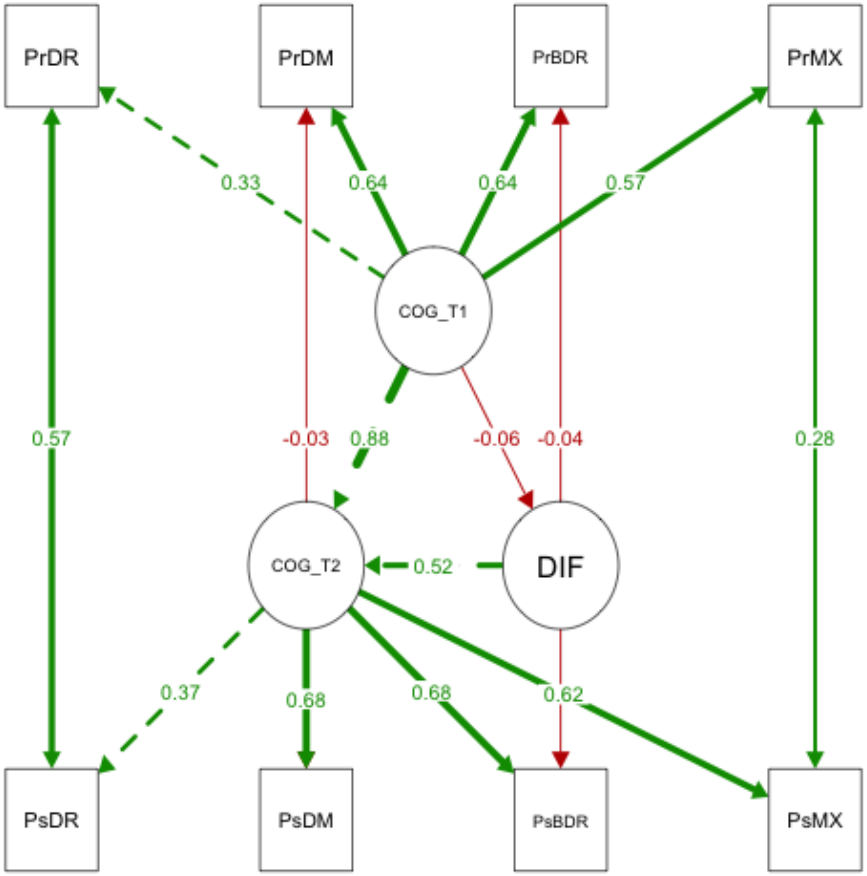
Note. * denotes statistical significance at level of .05.

Latent Change Score Modelling

As another multivariate statistical approach, Latent Change Score Modelling (LCSM) has also been adopted to model changes over time in the cognitive training research literature (Karbach et al., 2017; Schmiedek et al. 2010). To characterize the extent to which LCSMs can be used to understand training-induced effects, I fit a multiple indicator univariate latent change score model to the combined training dataset using R's Lavaan package and codes adapted from Kievit et al, 2017. I specified a model where all four cognitive measures load onto one latent variable "COG" (Figure A.7).

As discussed in the main text, LCSMs assume measurement invariance between the time points of assessments, namely, the latent variables are constrained to have the same unstandardized factor loadings and intercepts over time. I tested this assumption by comparing different levels of invariance (i.e. configural, metric, scalar and strict) using a chi-square difference test (Widaman et al. 2010). The model failed to achieve metric measurement invariance (i.e. fixed factor loadings across time) when compared to a model that assumed configural measurement invariance: $\Delta\chi2\ (3) = 8.125$, $p = 0.04$, suggesting that the relationships between the observed and the latent variable at Pre- and Post- training are not equivalent. To identify the same latent construct longitudinally, metric or strict invariance must hold across times of measurement; configural and weak invariance are insufficient (Widaman et al. 2010). Therefore, the result raised questions about the suitability of LCSMs within the scope of these combined training studies and necessitates the importance of having statistical alternatives that are also multivariate in nature.

Figure A.7. Latent Change Score Model on pre- and post-training data.



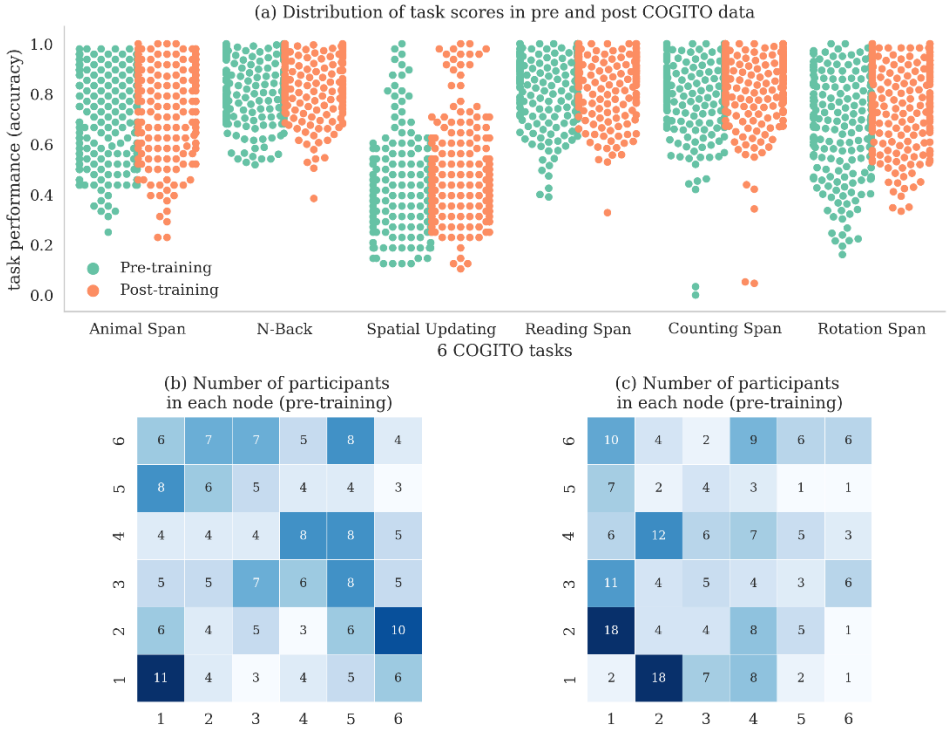
Note. COG_T1: latent factor at pre-training. COG_T2: latent factor at post-training. DIF: latent change score. There is no significant correlation between T1 and latent change (i.e. baseline score does not predict magnitude of improvement). Standardised parameters are displayed.

Analysis of COGITO data

To explore the scope of application of the SOM, an analysis using the same pipeline described in the main text was also applied to a separate dataset, namely the COGITO study from Schmiedek et al. (2010). COGITO data consisted of 204 participants who have completed an average of 100 hours of extensive training on working memory, processing speed and episodic memory (see the original published work for more details on training procedures and outcome measures). Six WM transfer measures including 3 updating and 3 complex span tasks were included in the initial training of SOM models for pre- and post-assessments, respectively. Individual scores were shown in Figure A8a, clearly revealing a ceiling effect on most tasks with the exception of Spatial updating. Similarly, it can be observed that on tasks such as n-back and Counting span, weight values did not vary much across model nodes in both times of assessment. This means that the performance on these tasks were invariant to the differential profiles that might exist on the other tasks. A contrasting example would be the result shown in Figure 2.2a, where relatively clear gradients of SOM weights for the CALM/ACE data existed. Indeed, the authors of COGITO study also discussed the existence of ceiling effect and its potential implication for data analysis.

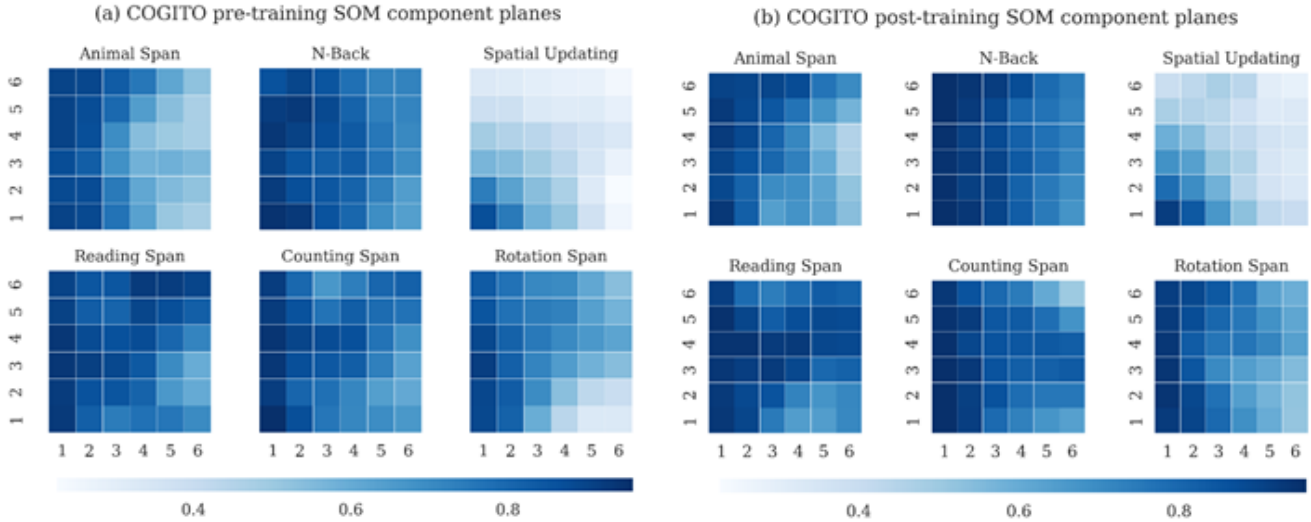
As such, I felt that these ceiling effects would constrain the ability to identify changes in task relationships or individual differences in training profiles and so decided against an in-depth analysis of the data, or its inclusion in the main body of the text.

Figure A.8. Overview on SOM analysis result of COGITO data (Schmiedek et al, 2010).



Note. (a) Task performance at pre- and post-training across 6 WM transfer tasks included the reported training study. Theoretical upper limit is at 1. (b) Distribution of participants in each node of a SOM fitted to pre-training data. (c) Distribution of participants at time of post-training in each node of the same SOM as in S8b. Note that large proportion of participant aggregated to the bottom left region which correspond to ceiling performance across all measures.

Figure A.9. Overview of SOM model weights trained on the COGITO (a) pre-training and (b) post-training data.



Note. It can be observed that on tasks such as n-back and Counting span (or even Animal span and Reading span), weight values did not vary much across model nodes in both times of assessment.

Comparison with the control group

Although the focus of this chapter is not on the efficacy of the training, I have included below a brief-surface level summary of the control data analysis for those interested. The SOM proved capable of predicting all four variables in the control dataset above chance. Furthermore, these predictions did not become significantly worse for any of the four variables (Unlike Dot Matrix in the training data), indicating that their relationships were represented somewhere in the model fit to the CALM/ACE data at both pre and post training (Table A.1).

Interestingly, the correlational analysis of task relationships as represented by a SOM model fit to the pre-training control data and the post-training control data respectively showed that some of these task relationships also change over time (Figure A.9), presumably due to practice effects/non-adaptive training effects, or test-retest effects. Specifically, the relationships between the Dot Matrix task and Forward Digit, Backward Digit, and Mr X were all altered significantly in the control groups data at post-training. Another explanation could be regression to the mean, as most of these children were screened for some format of low cognitive impairment, poor performers with relatively intact WM have been overrepresented and are more likely to regress upwards to their mean. Note that these changes are not in line with the adaptive training group, which speaks to the fact that the correlational trajectory of these tasks over time is likely, non-linear.

In a similar vein, the clustering analysis below, that involved assigning control participants to one of the four groups identified with K-means clustering on the CALM/ACE Model (Figure A.10) shows that even non-adaptive training can result in substantial gains, leading in this case to a re-assignment of sub-group. Again, these are subject to individual differences.

Table A.1. CALM/ACE-SOM prediction errors for the Pre- and Post-training control samples, and direct comparison of these prediction errors relative to one another.

		Forward Digit	Dot Matrix	Backward Digit Span	Mr.X
Pre-training (controls)	Prediction error (standard score)	11.35	13.07	10.57	13.73
	p	<0.001***	<0.001***	<0.001***	<0.01**
Post-training (controls)	Prediction error (standard score)	11.99	13.66	9.03	13.22
	p	<0.001***	<0.001***	<0.001***	<0.001***
Pre vs Post comparison (controls)	Difference in prediction error	0.64	0.59	-1.54	-0.15
	p	0.31	0.34	0.87	0.63

Note. Prediction error was defined as mean absolute difference between the predicted scores and true scores. p-values were derived from comparing the prediction errors against the corresponding chance level distributions. The chance levels were achieved by randomly shuffling the order of the predicted scores, then subtracting the true scores for 100 times within each cross-validation literation, to obtain a null distributions of mean absolute difference. Asterisks denote statistical significance at *p< .05, **p < .01 or ***p< .00.

Figure A.10. Between task correlations at Pre-Training-Control, Post-Training-Control, and the difference between the two; a comparison between SOM weights and raw data

Note. * denotes statistical significance at level of .05.

Appendix B – Supplementary methods and analyses to Chapter 3

Feature coding

Table B.1 shows which of the five features are shared across tasks and Table B.2 show the coding for task overlap.

Table B.1. Task feature coding

Features	Task					
	SSP	SN	SSW	DSP	DN	DSW
Spikiness	1	0	1	1	0	1
Number	0	1	1	0	1	1
Simultaneous	1	1	1	0	0	0
Delay	0	0	0	1	1	1
Switching	0	0	1	0	0	1

Note. Displays a binary representation of whether a feature is present on not in the task (1=present, 0=absent). Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW).

Table B.2. Predictor variable coding

	Task Pairs SSPT					β	p-value	BF_{10}
	SSP-SN	SSP-SSW	SSP-DSP	SSP-DN	SSP-DSW			
Total proportion	1/3	2/4	1/3	0/4	1/5	0.17	0.010	5.52
Spikiness shared	0	0.5	1	0	0.5	0.12	0.042	1.29
Correlation	0.35	0.25	0.59	0.36	0.19	0.08	0.121	0.50
	Task Pairs DSWT					β	p-value	BF_{10}
	DSW-SSP	DSW-SN	DSW-SSW	DSW-DSP	DSW-DN			
Total proportion	1/5	1/5	3/5	2/4	2/4	0.04	0.215	0.20
Spikiness shared	1	0	0.5	1	0	0.29	<0.001	228.85
Correlation	0.29	0.12	0.36	0.19	0.39	0.00	0.458	0.15

Note. For the total proportions, the numerator corresponds to the number of shared features between the training and assessment task and the denominator corresponds to the total number of unique features across both the training and assessment task. 'Spikiness shared' corresponds to the proportion of trials requiring a spikiness judgement in the assessment task. Finally, correlation corresponds to the pre-training correlation between the training and assessment task. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). P-values are group-wise holm-corrected.

Primary reaction time analyses

Pre-training performance

There was weak evidence for a difference between groups on the SSW task at pre training assessment (F(2,171) = 4.249, p=0.016, BF_{10}= 2.229, η_p^2=0.047). Post-hoc analyses provided strong evidence that the control group had lower reaction times than the DSWT group on the SSW task at pre-training assessment (t(113)=2.911, d=0.524 η_p^2=, p=0.012 , BF_{10}=6.418). There was no positive evidence in favour of either the null or alternative hypotheses for the SSW reaction times at pre-training assessment when comparing the control and SSPT groups (t(112)=1.374, d=0.236 η_p^2=, p=0.242, BF_{10}=0.405) nor when comparing the SSPT and DSWT groups (t(117)=1.559, d=0.332 η_p^2=, p=0.242, BF_{10}=0.846).

Table B.3. Assessment summary statistics for reaction time performance.

Task		Reaction Time (ms)						Paired t-test results				
Assessment	Training	Pre-training		Post-training		Difference (Post-Pre)						
		M	*SD*	*M*	*SD*	*M*	*SD*	*df*	*t*	*d*	BF_{10}	*p*
SSP	SSPT	903.92	106.86	860.94	125.55	-42.98	125.00	55	2.57	0.34	2.91	0.052
	DSWT	897.93	86.67	864.68	121.96	-33.25	111.30	59	2.31	0.29	1.65	0.040*
	Control	873.71	122.83	771.98	118.36	-101.73	114.53	52	6.46	0.88	>100	<0.001***
SN	SSPT	860.46	106.44	828.62	115.65	-31.84	110.47	57	2.19	0.28	1.31	0.096
	DSWT	861.37	112.82	826.77	115.36	-34.59	110.65	58	2.40	0.31	1.98	0.040*
	Control	822.92	96.30	755.60	119.36	-67.32	108.61	52	4.51	0.62	>100	<0.001***
SSW	SSPT	939.52	127.73	922.55	141.63	-16.96	113.50	58	1.14	0.14	0.26	0.512
	DSWT	979.03	109.98	929.16	140.96	-49.86	142.11	59	2.71	0.35	3.99	0.027*
	Control	903.94	172.42	849.72	168.92	-54.21	197.44	54	2.03	0.27	0.99	0.094
DSP	SSPT	665.18	102.18	623.42	77.15	-41.76	87.51	58	3.66	0.47	47.25	<0.001***
	DSWT	670.35	97.56	610.20	110.20	-60.15	93.48	58	4.94	0.64	>100	<0.001***
	Control	650.46	98.63	594.96	95.09	-55.50	85.90	53	4.74	0.64	>100	<0.001***
DN	SSPT	690.37	88.05	643.94	80.08	-46.43	86.53	57	4.08	0.53	>100	<0.001***
	DSWT	686.23	95.16	619.73	111.68	-66.50	102.89	55	4.83	0.64	>100	<0.001***
	Control	678.85	88.46	614.92	102.68	-63.92	80.99	53	5.79	0.78	>100	<0.001***
DSW	SSPT	721.01	136.47	706.85	119.29	-14.15	119.84	56	0.89	0.11	0.21	0.512
	DSWT	748.60	136.16	662.70	148.52	-85.89	117.80	57	5.55	0.72	>100	<0.001***
	Control	686.64	137.63	655.60	148.60	-31.03	117.00	55	1.98	0.26	0.90	0.094

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Task-wise holm-corrected).

Assessment task gains

Six paired sample t-tests were conducted for each group to establish whether participants made significant improvements on each of the assessment tasks relative to their baseline performance. The results are shown alongside descriptive summary statistics in Table B3, for transparency and completeness, but should not be interpreted as evidence of transfer.

Transfer effects

To investigate whether the groups show differential transfer patterns, a series of ANCOVAs were conducted to establish group differences in post-training performance, whilst controlling for pre-training performance. The full results are shown alongside the corresponding descriptive statistics for each task and each group contrast in Table B4. The positive evidence for group differences is summarised below.

Simultaneous Spikiness Training vs Control

The Control group had lower RTs relative to the Simultaneous Spikiness Training group on the Simultaneous Spikiness ($p<0.001$, $BF_{10}=50.08$, $\eta_p^2=0.23$) and Simultaneous Number ($p=0.033$, $BF_{10}=4.31$, $\eta_p^2=0.05$) tasks.

Delayed Switching Training vs Control

The Control group had lower RTs relative to the Delayed Switching Training group on the Simultaneous Spikiness ($p<0.001$, $BF_{10}>100$, $\eta_p^2=0.12$) and Simultaneous Number ($p=0.033$, $BF_{10}=3.66$, $\eta_p^2=0.05$) tasks.

Simultaneous Spikiness Training vs Delayed Switching Training

The Delayed Switching Training group had lower RTs relative to the Simultaneous Spikiness Training group on the Delayed Switching task ($p=0.009$, $BF_{10}=10.73$, $\eta_p^2=0.37$).

Table B.4. Pairwise group ANCOVAs of post-training reaction times adjusted for baseline performance.

Task	Post-training reaction time difference (ms)	ANCOVA				
		df	*F*	*p*	BF_{10}	η_p^2
SSPT-Control						
SSP	73.28	(1,106)	12.53	<0.001***	50.08	0.23
SN	51.01	(1,108)	6.74	0.033*	4.31	0.05
SSW	56.07	(1,111)	4.58	0.105	1.61	0.04
DSP	21.16	(1,110)	2.51	0.348	0.62	0.02
DN	22.25	(1,109)	2.39	0.375	0.59	0.02
DSW	30.31	(1,110)	2.27	0.164	0.56	0.02
DSWT-Control						
SSP	78.69	(1,110)	15.82	<0.001***	>100	0.12
SN	48.96	(1,109)	6.38	0.033*	3.66	0.05
SSW	51.34	(1,112)	3.31	0.142	0.99	0.02
DSP	2.63	(1,110)	0.02	0.867	0.20	0.00
DN	-0.15	(1,107)	0.00	0.993	0.20	0.00
DSW	-37.58	(1,111)	3.07	0.164	0.73	0.02
SSPT- DSWT						
SSP	-7.16	(1,113)	0.12	0.729	0.20	0.00
SN	2.33	(1,114)	0.01	0.899	0.20	0.96
SSW	17.87	(1,116)	0.63	0.429	0.25	0.00
DSP	16.02	(1,115)	1.22	0.540	0.34	0.01
DN	22.04	(1,111)	1.93	0.375	0.50	0.24
DSW	60.96	(1,112)	9.29	0.009**	10.73	0.37

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected).

Speed-Accuracy aggregate analyses

I include here a brief exploration of the training related gains and potential transfer effects according to the number of Responses Correct per Second (RCS) aggregate metric.

Assessment task gains

Six paired sample t-tests were conducted for each group to establish whether participants made significant improvements on each of the assessment tasks relative to their baseline performance. The results are shown alongside descriptive summary statistics in Table B5, for transparency and completeness, but should not be interpreted as evidence of transfer.

Table B.5. Assessment summary statistics for RCS performance.

Tasks		RCS						Paired t-test results				
Assessment	Training	Pre-training		Post-training		Difference (Post-Pre)						
		M	*SD*	*M*	*SD*	*M*	*SD*	*df*	*t*	*d*	BF_{10}	*p*
SSP	SSPT	0.84	0.12	0.99	0.14	0.15	0.15	55	7.19	0.96	>100	<0.001***
	DSWT	0.85	0.10	0.94	0.12	0.09	0.11	59	6.01	0.77	>100	<0.001***
	Control	0.87	0.13	0.99	0.13	0.11	0.15	52	5.77	0.79	>100	<0.001***
SN	SSPT	0.93	0.12	0.97	0.14	0.04	0.14	57	1.72	0.22	0.57	0.089
	DSWT	0.94	0.13	0.96	0.13	0.02	0.14	58	0.78	0.10	0.19	0.435
	Control	0.96	0.14	1.03	0.17	0.07	0.16	52	3.27	0.45	15.93	0.002**
SSW	SSPT	0.71	0.10	0.80	0.13	0.09	0.12	58	5.65	0.73	>100	<0.001***
	DSWT	0.70	0.10	0.80	0.11	0.10	0.11	59	6.68	0.86	>100	<0.001***
	Control	0.72	0.18	0.82	0.15	0.10	0.21	54	3.44	0.46	24.99	0.002**
DSP	SSPT	1.03	0.18	1.13	0.18	0.10	0.15	58	5.38	0.70	>100	<0.001***
	DSWT	0.99	0.14	1.17	0.21	0.18	0.18	58	7.28	0.94	>100	<0.001***
	Control	1.00	0.18	1.10	0.16	0.10	0.16	53	4.36	0.59	363.98	<0.001***
DN	SSPT	1.07	0.15	1.18	0.16	0.11	0.17	57	4.82	0.63	>100	<0.001***
	DSWT	1.07	0.17	1.25	0.20	0.18	0.19	55	6.65	0.88	>100	<0.001***
	Control	1.08	0.17	1.23	0.21	0.15	0.20	53	5.48	0.74	>100	<0.001***
DSW	SSPT	0.88	0.15	0.94	0.14	0.06	0.15	56	3.11	0.41	10.58	0.006**
	DSWT	0.84	0.17	1.07	0.24	0.23	0.20	57	8.90	1.16	>100	<0.001***
	Control	0.89	0.18	1.00	0.23	0.11	0.20	55	4.01	0.53	>100	<0.001***

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Task-wise holm-corrected). Responses Correct per Second (RCS)-Number of correct responses per second, an aggregate of accuracy and reaction time.

Transfer effects

To investigate whether the groups show differential transfer patterns, a series of ANCOVAs were conducted to establish group differences in post-training performance, whilst controlling for pre-training performance. The full results are shown alongside the corresponding descriptive statistics for each task and each group contrast in Table B6. The positive evidence for group differences is summarised below

Table B.6. Pairwise group ANCOVAs of post-training RCS adjusted for baseline performance.

Group Contrast	Task	Post-training RCS difference	ANCOVA				
			df	*F*	*p*	BF_{10}	η_p^2
SSPT-Control	SSP	0.01	(1,106)	0.23	0.626	0.22	0.00
	SN	-0.05	(1,108)	3.93	0.098	1.21	0.03
	SSW	-0.01	(1,111)	0.52	0.471	0.25	0.00
	DSP	0.01	(1,110)	0.59	0.442	0.26	0.00
	DN	-0.04	(1,109)	1.73	0.380	0.44	0.01
	DSW	-0.04	(1,110)	2.33	0.129	0.56	0.02
DSWT-Control	SSP	-0.03	(1,110)	3.06	0.166	0.80	0.02
	SN	-0.06	(1,109)	6.84	0.030*	4.18	0.05
	SSW	-0.01	(1,112)	0.64	0.422	0.26	0.00
	DSP	0.07	(1,110)	5.86	0.051	2.62	0.05
	DN	0.02	(1,107)	0.45	0.502	0.25	0.00
	DSW	0.11	(1,111)	9.13	0.006**	9.85	0.07
SSPT- DSWT	SSP	0.05	(1,113)	5.25	0.069	2.02	0.04
	SN	0.01	(1,114)	0.40	0.523	0.23	0.00
	SSW	-0.00	(1,116)	0.01	0.917	0.19	0.00
	DSP	-0.05	(1,115)	3.71	0.112	0.99	0.03
	DN	-0.06	(1,111)	4.46	0.108	1.44	0.03
	DSW	-0.15	(1,112)	23.06	<0.001***	>100	0.17

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected). Responses Correct per Second (RCS)-Number of correct responses per second, an aggregate of accuracy and reaction time.

Simultaneous Spikiness Training vs Control

There was no positive evidence for any differences between the Simultaneous Spikiness Training group and the Control group with respect to RCS rate on any of the tasks.

Delayed Switching Training vs Control

The Control group had a higher RCS rate relative to the Delayed Switching Training group on the Simultaneous Number task (p=0.030, BF_{10}=4.18, η_p^2=0.05). In contrast, the

Delayed Switching Training group has a higher RCS rate relative to the Control group on the Delayed Switching task (p=0.006, BF_{10}=9.85, η_p^2=0.07).

Simultaneous Spikiness Training vs Delayed Switching Training

The Delayed Switching Training group had a higher RCS rate relative to the Simultaneous Spikiness Training group on the Delayed Switching task (p<0.001, BF_{10}>100, η_p^2=0.17).

Mixing Costs

Mixing costs were calculated as the difference in performance between the tasks involving switching between the two judgement types and their non-switching counterparts. Using one-way ANOVAs, I found no positive evidence for group differences of mixing costs at pre training. I conducted paired t-tests to look for differences in switch costs between pre and post-training and ANCOVAs to look for between group differences in mixing costs at post-training after adjusting for pre-training. These results are shown in full in tables B.7 and B.8.

Table B.7. Accuracy mixing cost statistics and improvements over time.

Task		Mixing-Costs						Paired t-test results				
Assessment	Training	Pre-training		Post-training		Difference (Post-Pre)						
		M	*SD*	*M*	*SD*	*M*	*SD*	*df*	*t*	*d*	BF_{10}	*p*
SSW	SSPT	11.62	8.42	8.98	6.74	-2.64	8.43	51	2.25	0.31	1.54	0.028*
	DSWT	9.67	6.96	6.17	5.29	-3.50	7.92	52	3.21	0.44	13.72	0.002**
	Control	13.00	7.68	7.39	6.81	-5.60	9.25	47	4.19	0.60	>100	<0.001***
DSW	SSPT	8.09	7.92	6.81	6.06	-1.27	9.07	51	1.01	0.14	0.24	0.31
	DSWT	7.65	7.24	3.93	5.12	-3.72	8.68	52	3.12	0.42	10.67	0.003**
	Control	9.00	8.93	6.22	6.08	-2.78	10.04	47	1.91	0.27	0.84	0.061

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001

Table B.8. Pairwise group ANCOVAs of post-training mixing costs adjusted for baseline performance.

Task	Group Contrasts in switch cost savings Post-Pre (RT)	ANCOVA				
		df	F	p	BF_{10}	η_p^2
		SSPT-Control				
SSW	4.42	(1,97)	0.04	0.829	0.21	0.00
DSW	-5.91	(1,97)	0.11	0.731	0.22	0.00
		DSWT-Control				
SSW	6.65	(1,98)	0.08	0.766	0.21	0.00
DSW	25.92	(1,98)	2.80	0.097	0.69	0.02
		SSPT- DSWT				
SSW	-8.47	(1,102)	0.25	0.612	0.23	0.00
DSW	-34.72	(1,102)	5.73	0.019*	2.43	0.05

Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001

Switching task analysis split by trial type

I investigated the primary transfer outcomes with respect to accuracy in the constituent tasks further by splitting performance into its constituent spikiness and enumeration judgments. See tables B.9 and B.10.

Table B.9. Assessment summary statistics for accuracy performance in the switching tasks split by judgement type.

Task			Accuracy (%)						Paired t-test results				
Assessment	Judgement	Training	Pre-training		Post-training		Difference (Post-Pre)						
			M	*SD*	*M*	*SD*	*M*	*SD*	*df*	*t*	*d*	BF_{10}	*p*
SSW	Spikiness	SSPT	66.79	11.37	75.36	10.86	8.56	9.15	58	7.18	0.93	>100	<0.001***
		DSWT	68.09	10.45	74.43	10.22	6.33	8.98	59	5.46	0.70	>100	<0.001***
		Control	63.34	9.60	69.31	10.28	5.97	9.80	54	4.51	0.60	>100	<0.001***
	Number	SSPT	64.86	9.16	69.91	8.17	5.05	8.74	58	4.43	0.57	>100	<0.001***
		DSWT	67.97	9.90	71.96	8.58	3.98	11.64	59	2.65	0.34	3.42	0.010*
		Control	62.88	10.92	66.91	8.43	4.02	11.70	54	2.55	0.34	2.78	0.028*
DSW	Spikiness	SSPT	59.55	8.38	62.09	8.76	2.53	9.72	56	1.96	0.26	0.87	0.054
		DSWT	59.14	9.01	64.68	8.20	5.54	9.03	57	4.67	0.61	>100	<0.001***
		Control	56.02	9.32	59.91	6.66	3.89	9.33	55	3.11	0.41	10.61	0.003**
	Number	SSPT	64.44	8.31	69.20	8.43	4.76	11.38	56	3.15	0.41	11.8	0.003**
		DSWT	62.65	7.38	71.89	9.64	9.23	10.44	57	6.73	0.88	>100	<0.001***
		Control	62.75	9.57	65.59	11.23	2.83	12.41	55	1.71	0.22	0.57	0.093

Note. Assessment task abbreviations: Simultaneous Switching (SSW); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Task-wise holm-corrected).

Table B.10. Group contrasts of post-training accuracy performance split by judgement type adjusted for pre-training performance

Task	Judgement	Post-training accuracy difference (%)	ANCOVA				
			df	*F*	*p*	BF_{10}	η_p^2
SSPT-Control							
SSW	Spikiness	3.97	(1,111)	6.02	0.048*	3.07	0.05
	Number	2.37	(1,111)	2.68	0.208	0.68	0.02
DSW	Spikiness	1.08	(1,110)	0.60	0.440	0.27	0.00
	Number	3.22	(1,110)	3.06	0.146	0.79	0.02
DSWT-Control							
SSW	Spikiness	2.34	(1,112)	2.09	0.302	0.55	0.01
	Number	4.00	(1,112)	6.33	0.039*	3.79	0.05
DSW	Spikiness	3.73	(1,111)	8.18	0.015*	8.19	0.06
	Number	6.33	(1,111)	11.27	0.003**	27.25	0.09
SSPT- DSWT							
SSW	Spikiness	1.73	(1,116)	1.35	0.302	0.35	0.01
	Number	-1.10	(1,116)	0.56	0.543	0.25	0.00
DSW	Spikiness	-2.75	(1,112)	3.59	0.122	0.99	0.03
	Number	-3.04	(1,112)	3.27	0.146	0.82	0.02

Note. Assessment task abbreviations: Simultaneous Switching (SSW); Delayed Switching (DSW). Training group abbreviations: Simultaneous Spikiness Training (SSPT); Delayed Switching Training (DSWT). *p < .05. **p < .01. ***p < .001 (Group-wise holm-corrected).

Comparison of group contrast effect sizes

I compared the positive effect sizes for the ANCOVA group contrasts in post-training accuracy performance to see if their magnitude reliably differed from one another. To do this I used a bootstrapping procedure that involved fitting the same models to bootstrapped distributions of the variables of interest 2000 times to form a distribution of effect size estimates and allowing us produce p values of the difference.

Simultaneous Spikiness Training vs Control

The group contrast effect size was significantly larger for the Simultaneous Spikiness task than the Delayed Spikiness (η_p^2=0.25> η_p^2=0.06, p=0.002).

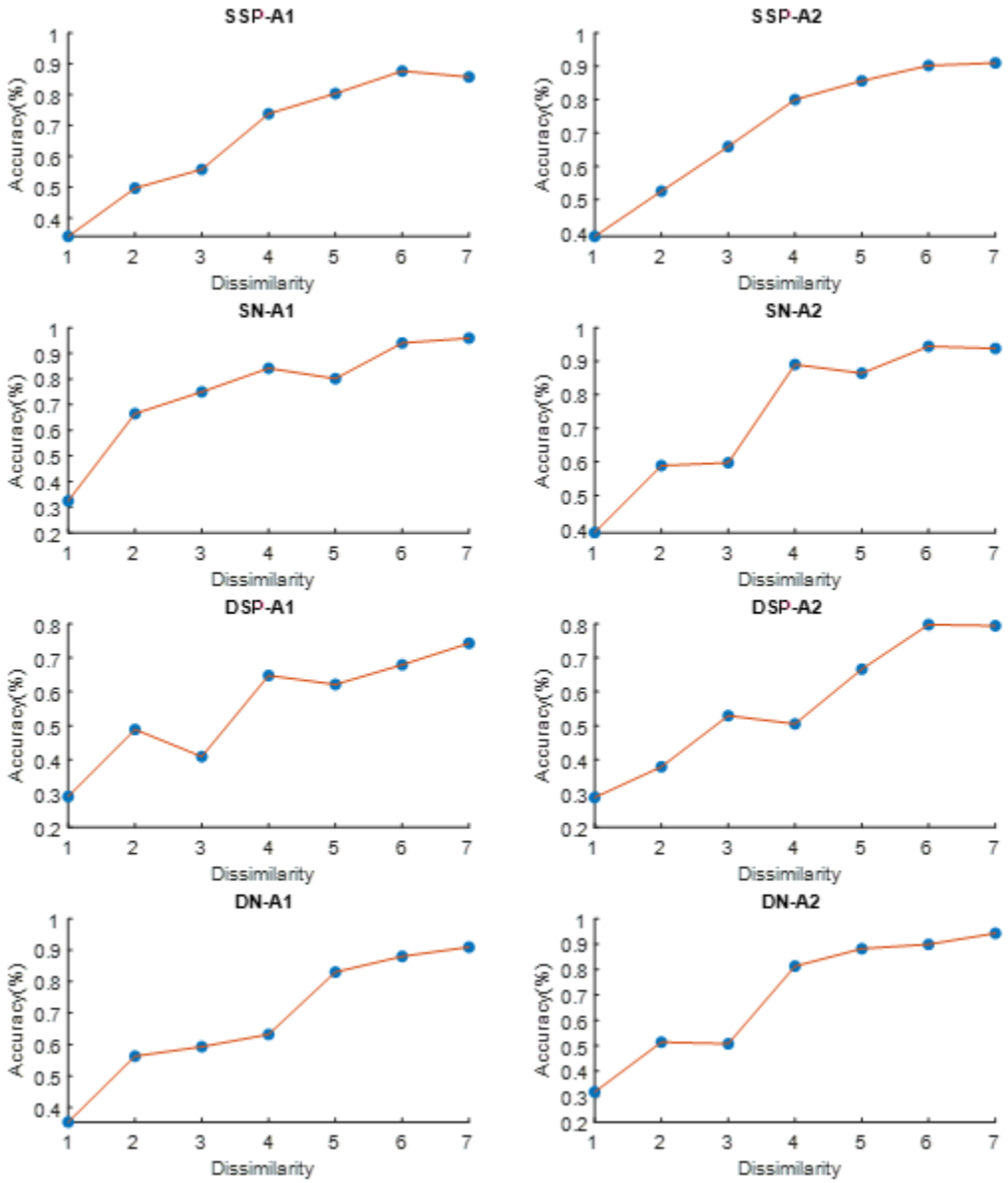
Delayed Switching Training vs Control

The group contrast effect was not significantly larger for the Delayed Switching task than the Simultaneous Spikiness task (η_p^2=0.12>η_p^2=0.08, p=0.114) or the Delayed Spikiness task (η_p^2=0.12>η_p^2=0.08, p=0.078). The group contrast effect was not significantly larger for the Simultaneous Spikiness task than the Delayed Spikiness Task (η_p^2=0.08>η_p^2=0.07, p=0.356).

Task Sensitivity

To get a rough idea of how sensitive our measures were before and after training I plotted task accuracy as a function of difficulty, i.e. dissimilarity (Figure B.1).

Figure B.1. Task accuracy as a function of dissimilarity



Note. Assessment task abbreviations: Simultaneous Spikiness (SSP); Simultaneous Number (SN); Simultaneous Switching (SSW); Delayed Spikiness (DSP); Delayed Number (DN); Delayed Switching (DSW). Pre-training assessment (A1) and post-training assessment (A2).

Appendix C – Supplementary methods and analyses to Chapter 4

Accuracy and reaction time summary statistics broken down by different factor combinations

The following pages contain summary statistics for Accuracy and Reaction time broken down by various factor combinations.

Table C.1. Descriptive statistics for the digit-span task pre and post.

Task	Training group	Span length					
		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD
Digit-Span	Ori-CDT	7.00	1.56	7.07	1.68	0.07	1.53
	Col-CDT	6.85	1.40	7.51	1.56	0.66	1.42
	Dual-CDT	6.41	1.24	7.02	1.45	0.60	1.24
	Digit-Span	6.83	1.57	9.83	2.31	3.00	2.05

Table C.2. Summary statistics for change detection accuracy performance across set-sizes pre and post, split by cue-type.

Task	Training group	Accuracy (%)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Ori-CDT	67.25	8.28	69.69	6.40	2.44	7.25	69.97	10.31	70.98	7.21	1.00	8.74	64.52	7.63	68.4	6.78	3.87	7.94
	Col-CDT	65.57	8.32	66.48	7.94	0.91	7.01	68.44	10.5	67.48	8.79	-0.97	8.61	62.70	7.66	65.47	8.02	2.78	8.08
	Dual-CDT	66.08	7.27	68.94	6.9	2.86	7.18	70.22	9.47	70.26	8.43	0.04	9.60	61.95	6.17	67.63	6.87	5.68	6.97
	Digit-Span	68.61	8.01	67.69	6.74	-0.92	6.21	73.01	10.04	69.65	8.23	-3.36	7.71	64.22	7.33	65.73	6.58	1.51	6.67
Col-CDT	Ori-CDT	64.67	8.15	67.35	7.37	2.69	8.17	66.38	9.22	71.06	7.50	4.68	10.25	62.96	8.32	63.65	8.71	0.69	7.80
	Col-CDT	64.85	8.83	75.57	9.68	10.72	10.93	67.07	9.80	78.70	9.59	11.63	11.11	62.63	8.99	72.44	10.34	9.81	11.63
	Dual-CDT	65.5	7.06	71.88	8.68	6.38	7.72	69.00	8.69	75.11	9.46	6.11	9.07	62.00	7.72	68.65	9.24	6.64	9.78
	Digit-Span	66.07	7.43	69.67	7.14	3.59	7.31	67.85	8.38	72.71	8.51	4.86	9.22	64.29	7.54	66.62	7.03	2.32	7.37
Dual-Ori-CDT	Ori-CDT	62.38	7.30	69.33	6.56	6.96	5.29	65.17	9.07	72.46	7.81	7.30	7.03	59.59	7.31	66.20	6.63	6.61	6.98
	Col-CDT	62.56	7.94	67.86	7.8	5.31	5.82	65.33	9.81	71.09	8.97	5.76	7.85	59.78	7.52	64.63	7.88	4.85	6.07
	Dual-CDT	63.40	8.13	69.64	6.54	6.24	7.85	67.12	9.71	73.14	8.05	6.02	9.48	59.69	7.38	66.15	6.62	6.46	8.23
	Digit-Span	65.24	7.80	66.80	7.88	1.57	6.72	69.19	10.26	71.19	9.96	2.00	8.73	61.28	6.50	62.42	7.23	1.14	7.62
Dual-Col-CDT	Ori-CDT	61.41	7.38	63.97	8.06	2.56	7.31	63.14	9.05	67.11	9.11	3.97	9.01	59.68	7.33	60.82	8.34	1.15	8.74
	Col-CDT	61.50	8.42	70.29	9.58	8.80	7.63	63.87	9.73	74.21	9.52	10.34	8.69	59.12	8.09	66.37	10.39	7.25	8.67
	Dual-CDT	61.95	7.77	70.01	7.83	8.06	7.25	64.90	9.02	74.23	9.17	9.33	9.32	58.99	7.57	65.79	7.71	6.80	6.97
	Digit-Span	62.21	7.68	65.51	8.58	3.30	7.12	64.56	10.06	68.37	10.13	3.81	9.10	59.85	7.01	62.65	7.89	2.80	7.89

Table C.3. Summary statistics for change detection reaction time performance across set sizes pre and post, split by cue-type.

Task	Training group	Reaction time (ms)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Ori-CDT	1198	359	709	146	-489	342	1089	363	701	144	-388	343	1307	364	717	150	-590	351
	Col-CDT	1117	367	963	193	-155	346	1006	352	958	199	-48	339	1229	388	967	191	-261	362
	Dual-CDT	1132	391	780	167	-353	349	1038	375	779	163	-258	338	1227	413	780	173	-447	369
	Digit-Span	1223	373	1030	293	-193	219	1102	371	1024	295	-77	212	1344	384	1035	294	-308	239
Col-CDT	Ori-CDT	1839	472	1345	375	-494	496	1759	442	1283	350	-476	472	1920	507	1407	406	-513	526
	Col-CDT	1781	456	1339	238	-442	435	1702	437	1251	217	-451	429	1861	483	1427	270	-434	453
	Dual-CDT	1633	455	1246	288	-386	465	1564	431	1183	281	-381	451	1702	489	1310	301	-392	490
	Digit-Span	1834	529	1540	511	-293	433	1752	510	1450	476	-303	393	1915	553	1631	552	-284	481
Dual-Ori-CDT	Ori-CDT	930	268	638	182	-293	195	869	269	605	177	-265	204	991	271	670	190	-321	191
	Col-CDT	1018	347	842	234	-177	245	961	345	791	244	-170	254	1075	358	893	232	-183	245
	Dual-CDT	918	295	610	193	-309	252	862	297	571	182	-292	250	975	300	649	210	-326	266
	Digit-Span	929	353	760	259	-168	190	856	344	709	249	-147	193	1001	368	812	274	-189	202
Dual-Col-CDT	Ori-CDT	1442	425	1244	338	-198	354	1419	436	1230	333	-189	371	1465	420	1258	348	-207	348
	Col-CDT	1580	448	1311	294	-269	344	1567	437	1272	277	-295	345	1593	466	1350	318	-243	355
	Dual-CDT	1504	466	1094	309	-410	430	1480	467	1067	299	-413	418	1528	469	1121	328	-407	454
	Digit-Span	1494	507	1355	517	-139	262	1451	482	1325	494	-126	265	1537	540	1384	546	-153	278

Table C.4. Summary statistics for change detection accuracy performance across cue-type pre and post, split by set-size.

Task	Training group	Accuracy (%)																	
		Set-size 2						Set-Size 4						Set-size 8					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Ori-CDT	77.78	10.74	84.64	6.78	6.86	10.97	66.67	10.39	65.36	8.04	-1.31	9.54	57.30	8.11	59.07	8.03	1.77	9.05
	Col-CDT	75.63	11.77	79.86	10.51	4.23	10.48	65.18	9.72	62.57	8.71	-2.61	9.57	55.89	6.85	56.99	7.85	1.1	7.43
	Dual-CDT	76.10	10.5	82.74	9.53	6.65	10.04	65.89	8.80	65.28	7.74	-0.61	10.90	56.26	6.73	58.8	8.01	2.54	8.04
	Digit-Span	80.38	10.66	82.23	8.49	1.86	9.07	67.65	10.12	63.83	9.25	-3.83	9.16	57.80	7.29	57.01	6.47	-0.80	7.20
Col-CDT	Ori-CDT	73.81	12.45	75.48	10.38	1.67	11.48	64.48	9.89	68.77	9.28	4.29	11.13	55.71	8.35	57.82	8.28	2.10	10.37
	Col-CDT	71.83	13.80	84.92	9.74	13.09	13.75	65.89	10.99	76.18	10.61	10.29	13.69	56.83	8.71	65.61	12.19	8.78	14.26
	Dual-CDT	71.87	10.89	81.54	10.00	9.67	11.92	67.28	9.34	73.33	9.91	6.06	9.98	57.36	7.39	60.75	10.13	3.40	10.43
	Digit-Span	74.09	11.27	79.60	9.73	5.51	12.40	67.54	9.13	69.20	9.73	1.67	10.71	56.59	6.51	60.19	7.92	3.60	8.93
Dual-Ori-CDT	Ori-CDT	70.44	9.95	77.91	7.38	7.47	7.96	61.67	9.24	69.54	7.56	7.87	8.31	55.03	6.43	60.55	7.85	5.52	6.96
	Col-CDT	69.55	10.97	76.76	9.62	7.22	9.90	62.74	9.69	67.95	9.85	5.22	6.32	55.39	7.74	58.88	6.80	3.49	7.72
	Dual-CDT	70.49	12.12	78.71	7.83	8.21	11.53	63.82	8.92	70.16	8.64	6.33	8.34	55.89	7.01	60.06	7.22	4.17	9.85
	Digit-Span	74.46	9.25	77.81	9.92	3.35	8.56	62.94	10.19	65.15	9.71	2.21	9.00	58.30	7.66	57.45	7.63	-0.85	8.37
Dual-Col-CDT	Ori-CDT	67.39	9.62	71.03	10.35	3.64	8.96	60.91	9.88	63.1	9.41	2.18	10.75	55.92	6.65	57.77	7.74	1.85	8.22
	Col-CDT	68.50	11.77	78.59	10.85	10.09	10.34	60.60	10.21	71.34	12.12	10.74	9.89	55.39	7.16	60.94	8.81	5.56	9.02
	Dual-CDT	69.75	10.51	80.00	9.23	10.25	10.04	69.75	60.40	9.12	69.8	10.13	9.40	55.69	7.94	60.23	7.96	4.54	9.01
	Digit-Span	69.51	11.19	74.12	11.11	4.61	11.19	61.90	10.16	64.43	11.10	2.53	10.83	55.21	5.89	57.99	7.36	2.78	7.02

Table C.5. Summary statistics for change detection reaction time performance across cue-type pre and post, split by set-size.

Task	Training group	Reaction time (ms)																	
		Set-size 2						Set-Size 4						Set-size 8					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Ori-CDT	1250	346	668	135	-581	346	1237	390	730	155	-508	376	1107	416	729	167	-379	372
	Col-CDT	1147	416	941	192	-206	413	1141	375	990	203	-151	355	1064	366	957	237	-107	317
	Dual-CDT	1151	407	748	162	-402	366	1143	409	788	177	-354	368	1104	384	803	183	-302	346
	Digit-Span	1244	365	1039	289	-204	236	1236	369	1057	299	-179	236	1189	449	993	327	-196	260
Col-CDT	Ori-CDT	2068	422	1472	356	-597	483	1836	507	1352	378	-483	534	1615	558	1212	425	-404	533
	Col-CDT	1942	476	1349	249	-594	444	1791	476	1346	248	-444	487	1613	524	1321	308	-292	444
	Dual-CDT	1788	458	1279	263	-509	456	1631	470	1249	286	-382	490	1480	505	1212	337	-268	511
	Digit-Span	2012	497	1657	483	-354	432	1836	523	1553	519	-282	450	1654	637	1411	580	-244	473
Dual-Ori-CDT	Ori-CDT	1007	273	634	183	-373	215	931	287	637	180	-294	209	853	286	642	199	-211	212
	Col-CDT	1062	352	854	273	-208	263	1027	344	843	224	-185	253	966	391	828	243	-138	272
	Dual-CDT	958	323	605	179	-353	278	915	295	610	203	-305	260	882	295	614	209	-268	252
	Digit-Span	996	371	801	275	-195	210	922	346	759	261	-163	197	868	386	721	272	-147	212
Dual-Col-CDT	Ori-CDT	1667	438	1363	326	-304	387	1416	441	1231	329	-186	353	1243	457	1138	380	-104	393
	Col-CDT	1754	453	1384	308	-370	331	1569	463	1310	309	-259	374	1418	535	1239	371	-178	398
	Dual-CDT	1653	468	1148	284	-505	420	1498	485	1089	321	-409	464	1361	489	1044	350	-317	469
	Digit-Span	1675	490	1517	497	-158	333	1466	530	1340	528	-126	282	1342	577	1208	587	-134	252

Table C.6. Summary statistics for the orientation-CDT accuracy performance split by group, set-size, and cue-type.

Training group	Set-size	Accuracy (%)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	67.25	8.28	69.69	6.40	2.44	7.25	69.97	10.31	70.98	7.21	1.00	8.74	64.52	7.63	68.4	6.78	3.87	7.94
	2	77.78	10.74	84.64	6.78	6.86	10.97	79.44	12.86	87.22	7.86	7.78	12.96	76.11	11.47	82.06	7.51	5.95	12.66
	4	66.67	10.39	65.36	8.04	-1.31	9.54	68.02	14.24	62.50	10.39	-5.52	14.22	65.32	10.36	68.21	8.75	2.90	11.76
	8	57.30	8.11	59.07	8.03	1.77	9.05	62.46	11.29	63.21	10.71	0.75	12.14	52.14	9.82	54.92	9.17	2.78	10.87
Col-CDT	Total	65.57	8.32	66.48	7.94	0.91	7.01	68.44	10.5	67.48	8.79	-0.97	8.61	62.70	7.66	65.47	8.02	2.78	8.08
	2	75.63	11.77	79.86	10.51	4.23	10.48	77.32	14.89	81.91	10.36	4.59	12.39	73.94	11.77	77.8	12.26	3.86	12.37
	4	65.18	9.72	62.57	8.71	-2.61	9.57	67.56	11.28	60.84	10.41	-6.72	12.94	62.81	10.93	64.31	9.43	1.50	10.85
	8	55.89	6.85	56.99	7.85	1.1	7.43	60.45	11.8	59.68	10.61	-0.77	12.18	51.34	6.88	54.31	9.44	2.97	10.94
Dual-CDT	Total	66.08	7.27	68.94	6.9	2.86	7.18	70.22	9.47	70.26	8.43	0.04	9.60	61.95	6.17	67.63	6.87	5.68	6.97
	2	76.10	10.5	82.74	9.53	6.65	10.04	80.49	11.27	84.59	10.66	4.11	12.94	71.71	11.01	80.89	10.17	9.19	10.77
	4	65.89	8.80	65.28	7.74	-0.61	10.90	69.02	11.91	63.01	10.46	-6.02	14.3	62.76	8.5	67.56	7.99	4.80	10.25
	8	56.26	6.73	58.8	8.01	2.54	8.04	61.14	10.4	63.17	11.1	2.03	11.13	51.38	7.67	54.43	9.41	3.05	11.5
Digit-Span	Total	68.61	8.01	67.69	6.74	-0.92	6.21	73.01	10.04	69.65	8.23	-3.36	7.71	64.22	7.33	65.73	6.58	1.51	6.67
	2	80.38	10.66	82.23	8.49	1.86	9.07	83.03	13.9	85.3	8.14	2.27	11.7	77.73	9.72	79.17	10.86	1.44	11.24
	4	67.65	10.12	63.83	9.25	-3.83	9.16	72.88	10.87	61.97	12.79	-10.91	11.19	62.42	11.47	65.68	8.55	3.26	11.38
	8	57.80	7.29	57.01	6.47	-0.80	7.20	63.11	10.96	61.67	9.30	-1.44	10.77	52.5	8.78	52.35	8.52	-0.15	10.37

Table C.7. Summary statistics for the colour-CDT accuracy performance split by group, set-size, and cue-type.

Training group	Set-size	Accuracy (%)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	64.67	8.15	67.35	7.37	2.69	8.17	66.38	9.22	71.06	7.50	4.68	10.25	62.96	8.32	63.65	8.71	0.69	7.80
	2	73.81	12.45	75.48	10.38	1.67	11.48	73.97	14.75	79.21	10.33	5.24	13.34	73.65	12.92	71.75	12.5	-1.90	12.96
	4	64.48	9.89	68.77	9.28	4.29	11.13	67.46	11.04	72.46	10.12	5.00	14.32	61.51	12.03	65.08	12.54	3.57	14.25
	8	55.71	8.35	57.82	8.28	2.10	10.37	57.70	9.41	61.51	9.46	3.81	13.1	53.73	11.90	54.13	11.54	0.40	14.56
Col-CDT	Total	64.85	8.83	75.57	9.68	10.72	10.93	67.07	9.80	78.70	9.59	11.63	11.11	62.63	8.99	72.44	10.34	9.81	11.63
	2	71.83	13.80	84.92	9.74	13.09	13.75	72.28	13.73	86.67	9.07	14.39	13.45	71.38	15.37	83.17	12.20	11.79	16.48
	4	65.89	10.99	76.18	10.61	10.29	13.69	68.54	13.15	82.19	10.16	13.66	14.10	63.25	11.46	70.16	12.58	6.91	16.72
	8	56.83	8.71	65.61	12.19	8.78	14.26	60.41	11.79	67.24	14.16	6.83	17.59	53.25	9.44	63.98	12.00	10.73	13.65
Dual-CDT	Total	65.5	7.06	71.88	8.68	6.38	7.72	69.00	8.69	75.11	9.46	6.11	9.07	62.00	7.72	68.65	9.24	6.64	9.78
	2	71.87	10.89	81.54	10.00	9.67	11.92	73.09	12.68	82.44	11.16	9.35	13.71	70.65	12.07	80.65	11.11	10.00	13.10
	4	67.28	9.34	73.33	9.91	6.06	9.98	73.09	10.73	78.25	11.09	5.16	11.78	61.46	11.62	68.41	12.11	6.95	14.50
	8	57.36	7.39	60.75	10.13	3.40	10.43	60.81	10.16	64.63	12.91	3.82	13.71	53.9	8.88	56.87	11.33	2.97	13.75
Digit-Span	Total	66.07	7.43	69.67	7.14	3.59	7.31	67.85	8.38	72.71	8.51	4.86	9.22	64.29	7.54	66.62	7.03	2.32	7.37
	2	74.09	11.27	79.60	9.73	5.51	12.40	73.79	13.09	80.15	10.39	6.36	13.62	74.39	11.56	79.05	10.99	4.66	14.09
	4	67.54	9.13	69.20	9.73	1.67	10.71	70.68	11.06	74.51	13.07	3.83	14.45	64.39	9.80	63.90	10.01	-0.49	11.98
	8	56.59	6.51	60.19	7.92	3.60	8.93	59.09	8.84	63.49	10.92	4.39	13.29	54.09	8.80	56.89	11.17	2.80	13.37

Table C.8. Summary statistics for the Dual-Orientation-CDT accuracy performance split by group, set-size, and cue-type.

Training group	Set-size	Accuracy (%)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	62.38	7.30	69.33	6.56	6.96	5.29	65.17	9.07	72.46	7.81	7.30	7.03	59.59	7.31	66.20	6.63	6.61	6.98
	2	70.44	9.95	77.91	7.38	7.47	7.96	73.74	12.04	79.56	9.16	5.82	9.40	67.13	10.52	76.26	8.35	9.13	10.54
	4	61.67	9.24	69.54	7.56	7.87	8.31	66.14	12.83	75.26	9.76	9.13	13.25	57.21	10.23	63.82	8.06	6.61	9.62
	8	55.03	6.43	60.55	7.85	5.52	6.96	55.62	8.71	62.57	9.80	6.94	9.45	54.43	8.52	58.53	8.42	4.10	11.01
Col-CDT	Total	62.56	7.94	67.86	7.8	5.31	5.82	65.33	9.81	71.09	8.97	5.76	7.85	59.78	7.52	64.63	7.88	4.85	6.07
	2	69.55	10.97	76.76	9.62	7.22	9.90	72.15	12.83	79.13	11.08	6.98	11.08	66.94	11.58	74.39	10.51	7.45	13.06
	4	62.74	9.69	67.95	9.85	5.22	6.32	66.73	13.37	72.97	11.88	6.23	11.19	58.74	8.55	62.94	10.44	4.20	8.01
	8	55.39	7.74	58.88	6.80	3.49	7.72	57.11	9.82	61.18	9.03	4.06	12.39	53.66	10.46	56.57	8.53	2.91	11.02
Dual-CDT	Total	63.40	8.13	69.64	6.54	6.24	7.85	67.12	9.71	73.14	8.05	6.02	9.48	59.69	7.38	66.15	6.62	6.46	8.23
	2	70.49	12.12	78.71	7.83	8.21	11.53	74.05	12.84	81.3	8.52	7.25	11.75	66.94	14.08	76.12	10.00	9.18	14.58
	4	63.82	8.92	70.16	8.64	6.33	8.34	68.63	12.75	77.27	10.69	8.64	11.91	59.01	8.62	63.04	8.75	4.03	9.18
	8	55.89	7.01	60.06	7.22	4.17	9.85	58.67	10.08	60.84	10.33	2.17	12.86	53.12	8.88	59.28	8.66	6.17	13.26
Digit-Span	Total	65.24	7.80	66.80	7.88	1.57	6.72	69.19	10.26	71.19	9.96	2.00	8.73	61.28	6.50	62.42	7.23	1.14	7.62
	2	74.46	9.25	77.81	9.92	3.35	8.56	77.97	10.66	81.76	10.97	3.79	9.64	70.96	9.83	73.86	11.23	2.90	11.22
	4	62.94	10.19	65.15	9.71	2.21	9.00	68.88	13.85	72.28	13.72	3.41	13.76	57.01	9.27	58.02	9.53	1.01	10.50
	8	58.30	7.66	57.45	7.63	-0.85	8.37	60.73	10.49	59.53	10.52	-1.20	10.10	55.87	8.56	55.37	8.33	-0.50	11.31

Table C.9. Summary statistics for the Dual-Colour-CDT accuracy performance split by group, set-size, and cue-type.

Training group	Set-size	Accuracy (%)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	61.41	7.38	63.97	8.06	2.56	7.31	63.14	9.05	67.11	9.11	3.97	9.01	59.68	7.33	60.82	8.34	1.15	8.74
	2	67.39	9.62	71.03	10.35	3.64	8.96	68.78	10.53	74.54	11.14	5.75	11.76	66.01	11.09	67.53	12.45	1.52	12.21
	4	60.91	9.88	63.1	9.41	2.18	10.75	63.43	13.51	66.53	10.73	3.11	11.68	58.4	10.37	59.66	10.47	1.26	14.13
	8	55.92	6.65	57.77	7.74	1.85	8.22	57.21	9.82	60.25	10.97	3.04	12.88	54.63	7.38	55.29	8.34	0.66	10.52
Col-CDT	Total	61.50	8.42	70.29	9.58	8.80	7.63	63.87	9.73	74.21	9.52	10.34	8.69	59.12	8.09	66.37	10.39	7.25	8.67
	2	68.50	11.77	78.59	10.85	10.09	10.34	70.33	13.14	81.17	9.85	10.84	10.61	66.67	12.16	76.02	13.73	9.35	13.25
	4	60.60	10.21	71.34	12.12	10.74	9.89	64.36	12.61	76.29	11.98	11.92	12.18	56.84	10.49	66.40	13.79	9.55	12.31
	8	55.39	7.16	60.94	8.81	5.56	9.02	56.91	9.48	65.18	11.79	8.27	12.31	53.86	9.31	56.71	8.76	2.85	11.53
Dual-CDT	Total	61.95	7.77	70.01	7.83	8.06	7.25	64.90	9.02	74.23	9.17	9.33	9.32	58.99	7.57	65.79	7.71	6.80	6.97
	2	69.75	10.51	80.00	9.23	10.25	10.04	71.48	11.32	82.86	9.46	11.38	12.59	68.02	12.41	77.13	11.83	9.11	11.41
	4	60.40	9.12	69.8	10.13	9.40	9.10	64.50	10.73	74.29	12.6	9.79	12.85	56.3	10.63	65.31	11.84	9.01	12.76
	8	55.69	7.94	60.23	7.96	4.54	9.01	58.74	10.84	65.55	10.6	6.81	12.49	52.64	8.58	54.91	8.38	2.27	11.04
Digit-Span	Total	62.21	7.68	65.51	8.58	3.30	7.12	64.56	10.06	68.37	10.13	3.81	9.10	59.85	7.01	62.65	7.89	2.80	7.89
	2	69.51	11.19	74.12	11.11	4.61	11.19	70.58	13.01	76.01	12.48	5.43	14.50	68.43	12.13	72.22	11.46	3.79	12.67
	4	61.90	10.16	64.43	11.10	2.53	10.83	65.28	12.55	67.68	14.00	2.40	12.84	58.52	10.10	61.17	11.17	2.65	12.42
	8	55.21	5.89	57.99	7.36	2.78	7.02	57.83	9.72	61.43	8.99	3.60	9.63	52.59	7.20	54.55	9.07	1.96	11.20

Table C.10. Summary statistics for the orientation-CDT reaction time performance split by group, set-size, and cue-type.

Training group	Set-size	Accuracy (ms)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	1198	359	709	146	-489	342	1089	363	701	144	-388	343	1307	364	717	150	-590	351
	2	1250	346	668	135	-581	346	1178	372	661	135	-518	368	1322	342	676	140	-645	344
	4	1237	390	730	155	-508	376	1114	393	721	153	-393	377	1361	403	739	163	-622	392
	8	1107	416	729	167	-379	372	975	413	720	167	-255	365	1240	439	737	172	-503	405
Col-CDT	Total	1117	367	963	193	-155	346	1006	352	958	199	-48	339	1229	388	967	191	-261	362
	2	1147	416	941	192	-206	413	1064	409	938	206	-126	411	1231	431	944	184	-287	425
	4	1141	375	990	203	-151	355	1017	379	985	211	-32	365	1265	386	995	209	-271	369
	8	1064	366	957	237	-107	317	937	337	951	243	14	302	1190	412	963	239	-227	357
Dual-CDT	Total	1132	391	780	167	-353	349	1038	375	779	163	-258	338	1227	413	780	173	-447	369
	2	1151	407	748	162	-402	366	1093	424	752	157	-340	376	1208	401	745	177	-464	368
	4	1143	409	788	177	-354	368	1027	385	781	183	-246	349	1258	445	796	176	-463	402
	8	1104	384	803	183	-302	346	994	357	805	187	-189	336	1215	426	801	186	-414	375
Digit-Span	Total	1223	373	1030	293	-193	219	1102	371	1024	295	-77	212	1344	384	1035	294	-308	239
	2	1244	365	1039	289	-204	236	1136	377	1043	299	-93	248	1351	363	1035	295	-316	253
	4	1236	369	1057	299	-179	236	1110	375	1052	311	-58	221	1362	389	1063	299	-300	296
	8	1189	449	993	327	-196	260	1059	430	978	322	-81	278	1319	483	1008	338	-311	289

Table C.11. Summary statistics for the colour-CDT reaction time performance split by group, set-size, and cue-type.

Training group	Set-size	Reaction time (ms)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	1839	472	1345	375	-494	496	1759	442	1283	350	-476	472	1920	507	1407	406	-513	526
	2	2068	422	1472	356	-597	483	1977	400	1401	350	-577	457	2159	457	1542	380	-617	525
	4	1836	507	1352	378	-483	534	1816	495	1311	351	-505	522	1856	539	1394	419	-462	568
	8	1615	558	1212	425	-404	533	1484	533	1138	404	-346	524	1747	596	1286	456	-461	563
Col-CDT	Total	1781	456	1339	238	-442	435	1702	437	1251	217	-451	429	1861	483	1427	270	-434	453
	2	1942	476	1349	249	-594	444	1863	467	1275	259	-588	456	2022	503	1423	266	-599	459
	4	1791	476	1346	248	-444	487	1740	470	1255	216	-485	461	1841	498	1437	297	-403	536
	8	1613	524	1321	308	-292	444	1505	515	1222	300	-283	467	1721	559	1420	336	-301	456
Dual-CDT	Total	1633	455	1246	288	-386	465	1564	431	1183	281	-381	451	1702	489	1310	301	-392	490
	2	1788	458	1279	263	-509	456	1713	454	1210	263	-502	457	1862	482	1347	275	-516	482
	4	1631	470	1249	286	-382	490	1601	449	1199	281	-402	474	1660	517	1299	309	-361	534
	8	1480	505	1212	337	-268	511	1378	472	1139	328	-238	500	1583	556	1284	354	-299	544
Digit-Span	Total	1834	529	1540	511	-293	433	1752	510	1450	476	-303	393	1915	553	1631	552	-284	481
	2	2012	497	1657	483	-354	432	1908	472	1564	471	-344	418	2116	541	1751	511	-364	481
	4	1836	523	1553	519	-282	450	1801	513	1477	462	-324	413	1870	549	1630	589	-240	513
	8	1654	637	1411	580	-244	473	1549	630	1309	566	-239	440	1760	663	1512	605	-248	529

Table C.12. Summary statistics for the Dual-Orientation-CDT reaction time performance split by group, set-size, and cue-type.

Training group	Set-size	Reaction time (ms)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	930	268	638	182	-293	195	869	269	605	177	-265	204	991	271	670	190	-321	191
	2	1007	273	634	183	-373	215	941	274	603	183	-338	233	1072	286	666	189	-407	217
	4	931	287	637	180	-294	209	866	299	594	172	-273	230	995	285	679	192	-316	205
	8	853	286	642	199	-211	212	800	280	618	198	-183	211	906	297	666	207	-240	227
Col-CDT	Total	1018	347	842	234	-177	245	961	345	791	244	-170	254	1075	358	893	232	-183	245
	2	1062	352	854	273	-208	263	1006	357	809	293	-197	284	1117	358	899	261	-218	260
	4	1027	344	843	224	-185	253	957	359	770	232	-187	269	1098	349	915	237	-182	255
	8	966	391	828	243	-138	272	920	369	793	248	-128	274	1012	423	864	250	-148	289
Dual-CDT	Total	918	295	610	193	-309	252	862	297	571	182	-292	250	975	300	649	210	-326	266
	2	958	323	605	179	-353	278	898	336	566	168	-333	283	1018	319	644	198	-374	286
	4	915	295	610	203	-305	260	854	311	560	188	-294	265	977	297	660	228	-317	287
	8	882	295	614	209	-268	252	835	283	586	201	-249	249	929	319	642	221	-287	269
Digit-Span	Total	929	353	760	259	-168	190	856	344	709	249	-147	193	1001	368	812	274	-189	202
	2	996	371	801	275	-195	210	913	362	742	268	-171	227	1078	390	860	293	-218	222
	4	922	346	759	261	-163	197	833	339	696	248	-137	202	1012	363	822	288	-190	230
	8	868	386	721	272	-147	212	822	375	689	262	-134	208	913	406	754	293	-160	241

Table C.13. Summary statistics for the Dual-Colour-CDT reaction time performance split by group, set-size, and cue-type.

Training group	Set-size	Reaction time (ms)																	
		Across cue-type						Cue						No-cue					
		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference		Pre-training		Post-training		Difference	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ori-CDT	Total	1442	425	1244	338	-198	354	1419	436	1230	333	-189	371	1465	420	1258	348	-207	348
	2	1667	438	1363	326	-304	387	1632	448	1328	312	-304	388	1703	450	1398	350	-304	423
	4	1416	441	1231	329	-186	353	1391	451	1223	327	-167	370	1442	441	1238	349	-204	359
	8	1243	457	1138	380	-104	393	1235	470	1140	391	-95	436	1250	454	1137	379	-113	364
Col-CDT	Total	1580	448	1311	294	-269	344	1567	437	1272	277	-295	345	1593	466	1350	318	-243	355
	2	1754	453	1384	308	-370	331	1729	461	1330	281	-399	354	1779	463	1439	351	-340	349
	4	1569	463	1310	309	-259	374	1558	451	1270	296	-288	403	1580	496	1350	344	-230	377
	8	1418	535	1239	371	-178	398	1415	522	1218	366	-198	391	1420	560	1261	396	-159	442
Dual-CDT	Total	1504	466	1094	309	-410	430	1480	467	1067	299	-413	418	1528	469	1121	328	-407	454
	2	1653	468	1148	284	-505	420	1607	492	1107	287	-500	434	1698	454	1189	296	-510	435
	4	1498	485	1089	321	-409	464	1463	497	1061	314	-402	470	1534	489	1118	341	-416	486
	8	1361	489	1044	350	-317	469	1372	474	1033	325	-339	438	1351	519	1055	386	-295	517
Digit-Span	Total	1494	507	1355	517	-139	262	1451	482	1325	494	-126	265	1537	540	1384	546	-153	278
	2	1675	490	1517	497	-158	333	1616	474	1467	467	-149	341	1735	520	1568	538	-167	359
	4	1466	530	1340	528	-126	282	1442	501	1317	514	-124	286	1490	573	1363	558	-128	319
	8	1342	577	1208	587	-134	252	1296	553	1193	565	-103	259	1389	613	1224	619	-165	282

ANCOVA results and pot-hoc follow ups

Table C.14. ANCOVAs testing for main effects and interactions on accuracy

Task	Main effects and Interactions	ANCOVA			
		df	*F*	*p*	η_p^2
Ori-CDT	Group	(3,983)	5.703	<0.001***	0.017
	Set-size	(2,983)	234.960	<0.001***	0.323
	Cue-type	(1,983)	0.621	0.430	0.000
	Group*Set-size	(6,983)	0.311	0.931	0.001
	Group*Cue-type	(3,983)	0.211	0.889	0.000
	Set-size*Cue-type	(2,983)	35.231	<0.001***	0.066
	Group*Set-size*Cue-type	(6,983)	0.621	0.713	0.003
Col-CDT	Group	(3,983)	25.239	<0.001***	0.071
	Set-size	(2,983)	118.545	<0.001***	0.194
	Cue-type	(1,983)	57.572	<0.001***	0.055
	Group*Set-size	(6,983)	0.995	0.427	0.006
	Group*Cue-type	(3,983)	0.328	0.805	0.001
	Set-size*Cue-type	(2,983)	4.170	0.015*	0.008
	Group*Set-size*Cue-type	(6,983)	1.396	0.212	0.008
Dual-Ori-CDT	Group	(3,983)	8.876	<0.001***	0.026
	Set-size	(2,983)	128.751	<0.001***	0.207
	Cue-type	(1,983)	60.377	<0.001***	0.057
	Group*Set-size	(6,983)	0.592	0.736	0.003
	Group*Cue-type	(3,983)	0.484	0.693	0.001
	Set-size*Cue-type	(2,983)	13.807	<0.001***	0.027
	Group*Set-size*Cue-type	(6,983)	0.968	0.445	0.005
Dual-Col-CDT	Group	(3,983)	23.983	<0.001***	0.068
	Set-size	(2,983)	81.987	<0.001***	0.143
	Cue-type	(1,983)	60.745	<0.001***	0.058
	Group*Set-size	(6,983)	1.795	0.096	0.010
	Group*Cue-type	(3,983)	0.656	0.579	0.002
	Set-size*Cue-type	(2,983)	0.662	0.516	0.001
	Group*Set-size*Cue-type	(6,983)	0.398	0.880	0.002
Digit-Span	Group	(3,162)	31.661	<0.001***	0.369

Note. *p < .05. **p < .01. ***p < .001.

Table C.15. Cue-type comparisons of the adjusted whole task mean accuracy differences adjusted for baseline performance.

Task	Cue-type contrast	Post-training accuracy difference (%)	t-test			
			df	*t*	*p*	*d*
Ori-CDT	Cue-no-cue	0.04	334	0.788	0.430	0.032
Col-CDT	Cue-no-cue	0.53	334	7.587	<0.001***	0.378
Dual-Ori	Cue-no-cue	0.46	334	7.770	<0.001***	0.366
Dual-Col	Cue-no-cue	0.51	334	7.793	<0.001***	0.441

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

Table C.16. Pairwise set size comparisons of the adjusted whole task mean accuracy differences adjusted for baseline performance.

Task	Set-size contrast	Post-training accuracy difference (%)	t-test			
			df	*t*	*p*	*d*
Ori-CDT	Two-Four	14.48	334	18.902	<0.001***	1.423
	Two-Eight	17.67	334	19.814	<0.001***	0.307
	Four-Eight	3.19	334	4.243	<0.001***	1.703
Col-CDT	Two-Four	6.84	334	7.806	<0.001***	0.558
	Two-Eight	15.16	334	15.389	<0.001***	1.262
	Four-Eight	8.32	334	9.227	<0.001***	0.659
Dual-Ori	Two-Four	6.44	334	8.979	<0.001***	0.565
	Two-Eight	12.92	334	16.043	<0.001***	1.305
	Four-Eight	6.48	334	8.979	<0.001***	0.590
Dual-Col	Two-Four	5.63	334	6.778	<0.001***	0.485
	Two-Eight	11.39	334	12.798	<0.001***	0.994
	Four-Eight	5.76	334	7.095	<0.001***	0.485

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

Table C.17. ANCOVAs testing for main effects and interactions on reaction time

Task	Main effects and Interactions	ANCOVA			
		df	*F*	*p*	η_p^2
Ori-CDT	Group	(3,983)	150.37	<0.001***	0.314
	Set-size	(2,983)	5.740	<0.01**	0.011
	Cue-type	(1,983)	16.440	<0.001***	0.016
	Group*Set-size	(6,983)	1.886	0.080	0.001
	Group*Cue-type	(3,983)	0.046	0.986	0.000
	Set-size*Cue-type	(2,983)	0.110	0.895	0.000
	Group*Set-size*Cue-type	(6,983)	0.108	0.995	0.000
Col-CDT	Group	(3,983)	21.614	<0.001***	0.061
	Set-size	(2,983)	0.568	0.566	0.001
	Cue-type	(1,983)	19.387	<0.001***	0.019
	Group*Set-size	(6,983)	1.816	0.092	0.011
	Group*Cue-type	(3,983)	0.416	0.740	0.001
	Set-size*Cue-type	(2,983)	0.028	0.972	0.000
	Group*Set-size*Cue-type	(6,983)	0.042	0.999	0.000
Dual-Ori-CDT	Group	(3,983)	69.599	<0.001***	0.175
	Set-size	(2,983)	2.703	0.067	0.005
	Cue-type	(1,983)	7.045	<0.01**	0.007
	Group*Set-size	(6,983)	1.294	0.256	0.007
	Group*Cue-type	(3,983)	0.600	0.614	0.001
	Set-size*Cue-type	(2,983)	0.737	0.478	0.001
	Group*Set-size*Cue-type	(6,983)	0.062	0.999	0.000
Dual-Col-CDT	Group	(3,983)	32.744	<0.001***	0.090
	Set-size	(2,983)	0.175	0.839	0.000
	Cue-type	(1,983)	1.986	0.159	0.002
	Group*Set-size	(6,983)	1.539	0.161	0.009
	Group*Cue-type	(3,983)	0.453	0.715	0.001
	Set-size*Cue-type	(2,983)	0.285	0.751	0.000
	Group*Set-size*Cue-type	(6,983)	0.051	0.999	0.000

Note. *p < .05. **p < .01. ***p < .001.

Table C.18. Cue-type comparisons of the adjusted whole task mean reaction time differences adjusted for baseline performance.

Task	Cue-type contrast	Post-training reaction time difference (ms)	t-test			
			df	*t*	*p*	*d*
Ori-CDT	Cue-no-cue	-51.66	334	4.054	<0.001***	0.199
Col-CDT	Cue-no-cue	-99.45	334	4.403	<0.001***	0.240
Dual-Ori	Cue-no-cue	-29.18	334	2.654	<0.01**	0.116
Dual-Col	Cue-no-cue	-27.64	334	1.409	0.159	0.065

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

Table C.19. Pairwise set size comparisons of the adjusted whole task mean reaction time differences adjusted for baseline performance.

Task	Set-size contrast	Post-training reaction time difference (ms)	t-test			
			df	*t*	*p*	*d*
Ori-CDT	Two-Four	44.44	334	2.954	<0.01**	1.728
	Two-Eight	43.96	334	2.912	<0.01**	0.169
	Four-Eight	0.48	334	0.031	0.974	0.001
Col-CDT	Two-Four	2.94	334	0.106	1.00	0.886
	Two-Eight	27.44	334	0.964	1.00	0.651
	Four-Eight	24.50	334	0.886	1.00	0.057
Dual-Ori	Two-Four	15.14	334	1.141	0.460	0.059
	Two-Eight	31.08	334	2.325	0.060	0.121
	Four-Eight	15.94	334	1.200	0.460	0.059
Dual-Col	Two-Four	6.15	334	0.253	1.00	0.015
	Two-Eight	14.70	334	0.588	1.00	0.035
	Four-Eight	8.54	334	0.353	1.00	0.020

Note. *p < .05. **p < .01. ***p < .001 (holm-corrected).

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, *8*(10), 457-464.

Alloway, T. P. (2007). Automated working memory battery for children.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of experimental child psychology*, *106*(1), 20-29.

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? Child development, 77(6), 1698-1716.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbour nonparametric regression. The American Statistician, 46(3), 175-185.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*(2), 106-111.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological review*, *89*(4), 369.

Astle, D. E., Nobre, A. C., & Scerif, G. (2012). Attentional control constrains visual short-term memory: Insights from developmental and individual differences. *Quarterly journal of experimental psychology*, *65*(2), 277-294.

Astle, D. E., Summerfield, J., Griffin, I., & Nobre, A. C. (2012). Orienting attention to locations in mental representations. *Attention, Perception, & Psychophysics*, *74*(1), 146-162.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic bulletin & review*, *22*(2), 366-377.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological science*, *18*(7), 622-628.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47-89). Academic press.

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current opinion in neurobiology*, *25*, 20-24.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, *128*(4), 612.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851-854.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of vision*, *9*(10), 7-7.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622-1631.

Berry, A. S., Zanto, T. P., Clapp, W. C., Hardy, J. L., Delahunt, P. B., Mahncke, H. W., & Gazzaley, A. (2010). The influence of perceptual training on working memory in older adults. *PloS one*, *5*(7).

Bialystok, E., Craik, F. I., Craik, F. I., & Craik, S. S. F. I. (Eds.). (2006). *Lifespan cognition: Mechanisms of change*. Oxford University Press, USA.

Bjork, E. L., & Bjork, R. A. (2014). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 59–68). New York (NY, USA): Worth.

Borella, E., Carretti, B., Cantarella, A., Riboldi, F., Zavagnin, M., & De Beni, R. (2014). Benefits of training visuospatial working memory in young-old and old-old. Developmental Psychology, 50(3), 714–727. https://doi.org/10.1037/a0034293

Burgess, P. W. (2004). Theory and methodology in executive function research. In Methodology of frontal and executive function (pp. 87-121). Routledge.

Bürki, C. N., Ludwig, C., Chicherio, C., & de Ribaupierre, A. (2014). Individual differences in cognitive plasticity: an investigation of training curves in younger and older adults. Psychological Research, 78, 821-835.

Buschkuehl, M., Jaeggi, S. M., Mueller, S. T., Shah, P., & Jonides, J. (2017). Training change detection leads to substantial task-specific improvement. *Journal of Cognitive Enhancement*, *1*(4), 419-433.

Campitelli, G., & Gobet, F. (2011). Deliberate practice: Necessary but not sufficient. *Current directions in psychological science*, *20*(5), 280-285.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic bulletin & review*, *17*(2), 193-199.

Chen, S. Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease*, *9*(6), 1725.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in cognitive sciences*, *21*(2), 111-124.

Clark, D. A., Nuttall, A. K., & Bowles, R. P. (2018). Misspecification in Latent Change Score Models: Consequences for Parameter Estimation, Model Evaluation, and Predicting Change. Multivariate Behavioral Research, 53(2), 172-189.

Cole, M. W., Laurent, P., & Stocco, A. (2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(1), 1-22.

Covey, T. J., Shucard, J. L., & Shucard, D. W. (2019). Working memory training and perceptual discrimination training impact overlapping and distinct neurocognitive processes: Evidence from event-related potentials and transfer of training gains. *Cognition*, *182*, 50-72.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive psychology*, *51*(1), 42-100.

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & cognition*, *34*(8), 1754-1768.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. Am. Psychol. 12, 671–684.

Dahlin, E.et al.(2008) Plasticity of executive functioning in young and older adults: immediate training gains, transfer, and long-term maintenance. Psychol. Aging 23, 720–730

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1-12.

Diamond, A. (2012). Activities and programs that improve children's executive functions. *Current directions in psychological science*, *21*(5), 335-341.

Dimitrov, D. M. (2006). Comparing groups on latent variables: A structural equation modeling approach. Work, 26(4), 429-436.

Dörrenbächer, S., Müller, P. M., Tröger, J., & Kray, J. (2014). Dissociable effects of game elements on motivation and cognition in a task-switching training in middle childhood. *Frontiers in psychology*, *5*, 1275.

Dosher, B. A., & Lu, Z. L. (2009). Hebbian reweighting on stable representations in perceptual learning. *Learning & Perception*, *1*(1), 37-58.

Dosher, B., & Lu, Z. L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science*, *3*, 343-363.

Dreisbach, G., & Wenke, D. (2011). The shielding function of task sets and its relaxation during task switching. Journal of Experimental Psychology: Learning Memory and Cognition, 37, 1540–1546

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends in neurosciences, 23(10), 475-483.

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks? Memory & cognition, 42(6), 854-62.

Dunning, D. L., Holmes, J., & Gathercole, S. E. (2013). Does working memory training lead to generalized improvements in children with low working memory? A randomized controlled trial. Developmental Science, 16(6), 915-925.

Ericcson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. Science, 208(4448), 1181-1182.

Fahle, M. (2005). Perceptual learning: specificity versus generalization. *Current opinion in neurobiology*, *15*(2), 154-160.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive science*, *6*(3), 205-254.

Ferguson, G. A. (1956). On transfer and the abilities of man. Can. J. Psychol. 10, 121–131.

Fine, I., & Jacobs, R. A. (2002). Comparing perceptual learning across tasks: A review. *Journal of vision*, *2*(2), 5-5.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of vision*, *11*(12), 3-3.

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory?. *Journal of vision*, *10*(12), 27-27.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic bulletin & review*, *17*(5), 673-679.

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. British Journal of Educational Psychology, 70(2), 177-194.

Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. Journal of Memory and Language, 105, 19-42.

Gogtay, N. J., & Thatte, U. M. (2017). Principles of correlation analysis. *Journal of the Association of Physicians of India*, *65*(3), 78-81.

Gray, J. R., & Thompson, P. M. (2004). Neurobiology of intelligence: science and ethics. Nature Reviews Neuroscience, 5(6), 471.

Green, C. S., & Bavelier, D. (2008). Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychology and aging*, *23*(4), 692.

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of cognitive neuroscience*, *15*(8), 1176-1194.

Guye, S., & von Bastian, C. C. (2017). Working memory training in older adults: evidence for the absence of transfer. Psychology and Aging.

Guye, S., De Simoni, C., & von Bastian, C. C. (2017). Do individual differences predict change in cognitive training performance? A latent growth curve modeling approach. Journal of Cognitive Enhancement, 1(4), 374-393.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. Psychological methods, 20(1), 102.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. Psychological Science, 24(12), 2409-2419.

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. Trends in cognitive sciences, 19(6), 304-313.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166-1186.

Hertzog, C., Kramer, A. F., Wilson, R. S., & Lindenberger, U. (2008). Enrichment effects on adult cognitive development: can the functional capacity of older adults be preserved and enhanced?. Psychological science in the public interest, 9(1), 1-65.

Heuer, A., & Schubö, A. (2016). Feature-based and spatial attentional selection in visual working memory. *Memory & cognition*, *44*(4), 621-632.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. Developmental science, 12(4).

Holmes, J., Gathercole, S. E., Place, M., Dunning, D. L., Hilton, K. A., & Elliott, J. G. (2010). Working memory deficits can be overcome: Impacts of training and medication on working memory in children with ADHD. Applied Cognitive Psychology, 24(6), 827-836.

Holmes, J., Woolgar, F., Hampshire, A., & Gathercole, S. E. (2019). Are working memory training effects paradigm-specific?. *Frontiers in psychology*, *10*, 1103.

Holmes, J.et al.(2009) Working memory deficits can be overcome: impacts of training and medication on working memory in children with ADHD. Appl. Cogn. Psychol.12, 9–15

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & cognition*, *42*(3), 464-480.

Jeter, P. E., Dosher, B. A., Liu, S. H., & Lu, Z. L. (2010). Specificity of perceptual learning increases with increased training. *Vision research*, *50*(19), 1928-1940.

Jeter, P. E., Dosher, B. A., Petrov, A., & Lu, Z. L. (2009). Task precision at transfer determines specificity of perceptual learning. *Journal of vision*, *9*(3), 1-1.

Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Hahn, B., Leonard, C. J., ... & Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology*, *27*(2), 220.

Kandel, E. R. (2007). *In search of memory: The emergence of a new science of mind*. WW Norton & Company.

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental science*, *12*(6), 978-990.

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. Psychological Science, 25, 2027–2037.

Karbach, J., Könen, T., & Spengler, M. (2017). Who benefits the most? Individual differences in the transfer of executive control training across the lifespan. Journal of Cognitive Enhancement, 1(4), 394-405.

Katz, B., Shah, P., & Meyer, D. E. (2018). How to play 20 questions with nature and lose: Reflections on 100 years of brain-training research. *Proceedings of the National Academy of Sciences*, *115*(40), 9897-9904.

Kievit, R. A. (2020). Sensitive periods in cognitive development: A mutualistic perspective. *Current Opinion in Behavioral Sciences*, *36*, 144-149.

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A. L., de Mooij, S. M., Moutoussis, M., ... & Lindenberger, U. (2017). Developmental cognitive neuroscience

using Latent Change Score models: A tutorial and applications. Developmental cognitive neuroscience.

Kievit, R. A., Romeijn, J. W., Waldorp, L. J., Wicherts, J. M., Scholte, H. S., & Borsboom, D. (2011). Modeling mind and matter: reductionism and psychological measurement in cognitive neuroscience. *Psychological Inquiry*, *22*(2), 139-157.

Kliegl, R., Smith, J., & Baltes, P. B. (1990). On the locus and process of magnification of age differences during mnemonic training. Developmental Psychology, 26(6), 894–904. https://doi.org/10.1037/0012-1649.26.6.894

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in cognitive sciences*, *14*(7), 317-324.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., ... & Westerberg, H. (2005). Computerized training of working memory in children with ADHD-a randomized, controlled trial. Journal of the American Academy of Child & Adolescent Psychiatry, 44(2), 177-186.

Klingberg, T., Forssberg, H., and Westerberg H. (2002). Training of working memory in children with ADHD. *J. Clin. Exp. Neuropsychol.* 24, 781–791. doi: 10.1076/jcen.24.6.781.8395

Klingberg, T.et al.(2005) Computerized training of working memory inchildren with ADHD–a randomized, controlled trial.J. Am. Acad.Child Adolesc. Psychiatry44, 177–186

Kohonen (2013) Matlab Documentation-find this on my computer

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.

Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21(1-3), 1-6.

Li, S. C., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., & Lindenberger, U. (2008). Working memory plasticity in old age: practice gain, transfer, and maintenance. Psychology and aging, 23(4), 731.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis*. Lawrence Erlbaum Associates, Inc.

Lövdén, M., Brehmer, Y., Li, S.-C., & Lindenberger, U. (2012). Training-induced compensation versus magnification of individual differences in memory performance. Frontiers in Human Neuroscience, 141. https://doi.org/10.3389/fnhum.2012.00141.

Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., & Tucker-Drob, E. M. (2020). Education and cognitive functioning across the life span. *Psychological Science in the Public Interest*, *21*(1), 6-41.

Lu, Z. L., & Dosher, B. A. (2009). Mechanisms of perceptual learning. *Learning & perception*, *1*(1), 19-36.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, *17*(8), 391-400.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. *JMIR serious games*, *4*(2), e11.

Luria, R., & Vogel, E. K. (2011). Shape and color conjunction stimuli are represented as bound objects in visual working memory. *Neuropsychologia*, *49*(6), 1632-1639.

MacQueen, J., et al. (1967). Some methods for classi_cation and analysis of multivariate observations. In Proceedings of the _fth Berkeley symposium on mathematical statistics and probability, vol. 1, (pp. 281{297}. Oakland, CA, USA

Marshall, L., & Bays, P. M. (2013). Obligatory encoding of task-irrelevant features depletes working memory resources. *Journal of vision*, *13*(2), 21-21.

Maul, A., Irribarra, D. T., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, *79*, 311-320.

Melby-Lervag, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. Developmental psychology, 49(2), 270.

Melby-Lervåg, M., & Hulme, C. (2016). There is no convincing evidence that working memory training is effective: A reply to Au et al.(2014) and Karbach and Verhaeghen (2014). Psychonomic Bulletin & Review, 23(1), 324-330.

Melby-Lervag, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of “far transfer” evidence from a meta-analytic review. Perspectives on Psychological Science, 11(4), 512-534.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R.,& Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. American psychologist, 56(2), 128.

Minear, M. E., Shah, P., & Park, D. (2002). Age, task switching, and transfer of training. In *Poster presented at the Ninth Cognitive Aging Conference, Atlanta, GA*.

Minear, M., & Shah, P. (2008). Training and transfer effects in task switching. *Memory & cognition*, *36*(8), 1470-1483.

Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. Memory & cognition, 44(7), 1014-1037.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, *41*(1), 49-100.

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., ... & Seitz, A. R. (2017). The benefits and challenges of implementing motivational features to boost cognitive training outcome. *Journal of Cognitive Enhancement*, *1*(4), 491-507.

Moreau, D., Kirk, I. J., & Waldie, K. E. (2016). Seven pervasive statistical flaws in cognitive training interventions. Frontiers in human neuroscience, 10, 153.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, *1*(1981), 1-55.

Norris, D. G., Hall, J., & Gathercole, S. E. (2019). Can short-term memory be trained?. *Memory & cognition*, *47*(5), 1012-1023.

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science*, *21*(3), 164-169.

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. *Perception & psychophysics*, *64*(7), 1055-1067.

Parsons, B., Magill, T., Boucher, A., Zhang, M., Zogbo, K., Bérubé, S., ... & Faubert, J. (2016). Enhancing cognitive function using perceptual-cognitive training. *Clinical EEG and neuroscience*, *47*(1), 37-47.

Prins, N. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes toolbox. *Frontiers in psychology*, *9*, 1250.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3-25.

Protzko, J. (2017). Effects of cognitive training on the structure of intelligence. Psychonomic bulletin & review, 24(4), 1022-1031.

Reder, L., & Klatzky, R. L. (1994). *The effect of context on training: Is learning situated?* (No. CMU-CS-94-187). CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

Redick, T. S. (2019). The hype cycle of working memory training. *Current Directions in Psychological Science*, *28*(5), 423-429.

Rennie, J. P., Zhang, M., Hawkins, E., Bathelt, J., & Astle, D. E. (2020). Mapping differential responses to cognitive training using machine learning. *Developmental science*, *23*(4), e12868.

Rennie, J., Jones, J., & Astle, D. (2021). EXPRESS: Training-dependent transfer within a set of nested tasks. *Quarterly Journal of Experimental Psychology*, 1747021821993772.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. Journal of Experimental Psychology: General, 124, 207-231.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975-5979.

Sabah, K., Dolk, T., Meiran, N., & Dreisbach, G. (2019). When less is more: Costs and benefits of varied vs. fixed content and structure in short-term task switching training. Psychological research, 83, 1531-1542.

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in cognitive sciences*, *23*(1), 9-20.

Salthouse, T. A., & Davis, H. P. (2006). Organization of cognitive abilities and neuropsychological variables across the lifespan. *Developmental Review*, *26*(1), 31-54.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. Frontiers in aging neuroscience, 2, 27.

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *cortex*, *83*, 181-193.

Schneegans, S., & Bays, P. M. (2017). Neural architecture for feature binding in visual working memory. *Journal of Neuroscience*, *37*(14), 3913-3925.

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, *117*(34), 20959-20968.

Schweizer, K. (2007). Investigating the relationship of working memory tasks and fluid intelligence tests by means of the fixed-links model in considering the impurity problem. *Intelligence*, *35*(6), 591-604.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in medicine*, *25*(24), 4334-4344.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective?. Psychological bulletin, 138(4), 628.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work?. Psychological Science in the Public Interest, 17(3), 103-186.

Singley, M. K., & Anderson, J. R. (1985). The transfer of text-editing skill. *International Journal of Man-Machine Studies*, *22*(4), 403-423.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill* (No. 9). Harvard University Press.

Sligte, I. G., Scholte, H. S., & Lamme, V. A. (2008). Are there multiple visual short-term memory stores?. *PLOS one*, *3*(2), e1699.

Sligte, I. G., Vandenbroucke, A. R., Scholte, H. S., & Lamme, V. (2010). Detailed sensory memory, sloppy working memory. *Frontiers in psychology*, *1*, 175.

Smid, C. R., Karbach, J., & Steinbeis, N. (2020). Toward a science of effective cognitive training. *Current Directions in Psychological Science*, *29*(6), 531-537

Smoleń, T., Jastrzebski, J., Estrada, E., & Chuderski, A. (2018). Most evidence for the compensation account of cognitive training is unreliable. Memory & cognition, 46(8), 1315-1330.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P.L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences: advances in theory and research (pp. 13-59). New York, NY: W H Freeman/Times Books/Henry Holt & Co..

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839-1860.

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. Psychonomic Bulletin & Review, 24(4), 1077-1096.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., ... & Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence*, *41*(5), 638-663.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological review*, *120*(3), 439.

Taylor, R., & Bays, P. M. (2020). Theory of neural coding predicts an upper bound on estimates of memory variability. *Psychological review*.

Thorell, L.B.et al.(2009) Training and transfer effects of executivefunctions in preschool children.Dev. Sci.12, 106–113

Van Breukelen, G. J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of clinical epidemiology*, *59*(9), 920-925.

Van Dam, L. C., & Ernst, M. O. (2015). Mapping shape to visuomotor mapping: learning and generalisation of sensorimotor behaviour based on contextual information. *PLoS computational biology*, *11*(3).

Van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T., ... & Ly, A. (2019). The JASP guidelines for conducting and reporting a Bayesian analysis.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27 (1), 92–114.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436.

von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: a review. Psychological research, 78(6), 803-820.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*(1), 35-57.

Wechsler, D., Scales, P. I., & Index, V. C. (2012). Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition.

Wheeler, M., & Treisman, A. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology General*, 131 (1), 48–64.

Widaman, K. F., Ferrer, E. and Conger, R. D. (2010), Factorial Invariance Within Longitudinal Structural Equation Models: Measuring the Same Construct Across Time. Child Development Perspectives, 4: 10-18.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of vision*, *4*(12), 11-11.

Xu, Y. (2002). Limitations of object-based feature encoding in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, 28 (2), 458–468.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, *22*(2), 297-306.

Yin, H. (2008). The self-organizing maps: background, theories, extensions and applications. In Computational intelligence: A compendium (pp. 715-762). Springer, Berlin, Heidelberg.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233-235.

Zinke, K., Zeintl, M., Rose, N. S., Putzmann, J., Pydde, A., & Kliegel, M. (2014). Working memory training and transfer in older adults: effects of age, baseline performance, and training gains. Developmental Psychology, 50(1), 304–315. https://doi.org/10.1037/a0032982.